

# PHYSICS

for the Technician  
by L. S. Zhdanov







Л. С. Жданов

## УЧЕБНИК ПО ФИЗИКЕ

для средних специальных  
учебных заведений

*Издательство «Наука»*

*Москва*

# PHYSICS

**for the Technician**

**by L. S. Zhdanov**

Translated from the Russian  
by Mark Samokhvalov, Cand. Sc.

Mir Publishers  
Moscow



First published 1980  
Revised from the 1977 Russian edition  
*На английском языке*

- © Главная редакция физико-математической литературы  
издательства «Наука», 1975 г.
- © English translation, Mir Publishers, 1980

# CONTENTS

FOREWORD 15

## Introduction

### 1 PHYSICAL QUANTITIES AND THEIR MEASUREMENT 18

What is Physics? Physics and Technology. A Quantity and Its Measurement. Physical Quantities. Direct and Indirect Measurements. Measurement of Angles in Astronomy. Measuring Distances to Celestial Bodies by the Parallax Method. Units of Time and Their Relation to Earth's Motion. Units of Measurement from Formulae. The International System of Units. Treatment of Data. Combining Errors. Density of Substance.

## Heat and Molecular Physics

part one

### 2 THE FUNDAMENTALS OF KINETIC THEORY OF MATTER 36

First Principles of Kinetic Theory. Concept of Temperature. Diffusion. Forces of Molecular Interaction. Kinetic and Potential Energies of Molecules. Concept of Internal Energy. Probability of an Event. The Statistical Method.

### 3 KINETIC THEORY OF GASES 47

The Gaseous State. Brownian Motion. Measuring Molecular Speeds. Distribution of Molecular Speeds. Mass and Size of Mol-



ecules and Atoms. Avogadro and Loschmidt Numbers. Mean Free Path. Gaseous Pressure. Pressure Gauges. Kinetic Calculation of the Pressure. Vacuum.

#### **4 THE IDEAL GAS 61**

Properties of Ideal Gas. Change of Gaseous Pressure with Temperature at Constant Volume. Absolute Zero. Thermodynamic Temperature Scale. Relation of Temperature to Kinetic Energy of Gas Molecules.

#### **5 IDEAL-GAS EQUATION OF STATE 68**

Thermodynamic Properties. Combined Gas Law. Universal Gas Constant. The Ideal-Gas Law. Dependence of Root-Mean-Square Speed of Gas Molecules on Temperature. Isochoric Process. Isobaric Process. Isothermal Process. Internal Energy of Ideal Gas. Work Performed by Gas.

#### **6 INTERNAL ENERGY 80**

Internal Energy and the Surroundings. Heat Exchange. Types of Heat Exchange. Changing Internal Energy by Means of Work. Relation of Internal Energy to State of Matter.

#### **7 QUANTITY OF HEAT 86**

The Measurement of Heat. Changing Internal Energy by Heating or Cooling. Heat of Combustion. The Law of Heat Exchange.

#### **8 THE LAW OF CONSERVATION OF ENERGY. THE FIRST LAW OF THERMODYNAMICS 91**

Mechanical Equivalent of Heat. Conservation of Energy in Mechanics. The Law of Conservation of Energy. The First Law of Thermodynamics. Some Applications of the First Law of Thermodynamics. Adiabatic Process. Some Ideas on Stellar Structure.

#### **9 CHANGE OF STATE 101**

Vapourization and Condensation. Evaporation. Heat of Vapourization.

**10 PROPERTIES OF VAPOUR. BOILING 105**

Nonsaturated and Saturated Vapours. Properties of Saturated Vapour. Properties of Nonsaturated Vapour. The Boiling Process. Dependence of Boiling Temperature on External Pressure. Boiling Point. The Law of Heat Exchange for Vapourization and Condensation. Superheated Steam and Its Use in Technology. Critical State of Substance. Liquefaction of Gases.

**11 WATER VAPOUR IN THE ATMOSPHERE 118**

Humidity. Absolute and Relative Humidities. Measuring Humidity. The Atmosphere of Planets.

**12 THE LIQUID STATE 122**

What Is a Liquid? The Surface Layer of a Liquid. Surface Tension. Measuring Surface Tension. Wetting. The Shape of Liquid Surfaces. Capillarity. Viscosity. Newton's Law of Fluid Friction. Amorphous Substances.

**13 THE SOLID STATE 137**

What Is a Solid? Crystalline Anisotropy. Types of Crystals. Types of Deformation. Stress. Elasticity, Plasticity, Brittleness and Hardness. Hooke's Law. Energy of a Body Under Elastic Deformation.

**14 CHANGE OF STATE—II 151**

Fusion and Crystallization. Specific Heat of Fusion. Changes in Volume and Density During Fusion and Solidification. Pressure Dependence of Temperature of Fusion and Heat of Fusion. The Law of Heat Exchange for Fusion and Crystallization. Solutions and Alloys. Sublimation. Phase Diagrams. Triple Point.

**15 THERMAL EXPANSION 161**

Basic Facts About Thermal Expansion. Linear Expansion. Volume Expansion of Heated Bodies. Thermal Expansion of Solids. Thermal Expansion of Liquids. Thermal Expansion in Nature and Technology.



## part two

## Electricity and Magnetism

**16 THE FUNDAMENTALS OF THE ELECTRON THEORY OF ATOMIC STRUCTURE. COULOMB'S LAW 168**

Electrification of Bodies. The Concept of an Electric Charge. The Complex Nature of the Atomic Structure. Rutherford's Experiment and the Nuclear Idea. The Atomic Structure of Chemical Elements. Electrification by Contact. Interaction Between Electric Charges. Coulomb's Law. The Permittivity of a Medium. SI Units in Electricity. Gaussian Units in Electrostatics. The Electroscope.

**17 THE ELECTRIC FIELD 179**

Electric Field as a Special Form of Matter. The Electric Field Strength. Electric Field and Lines of Force. The Homogeneous Electric Field. Work Done by an Electric Field in Moving a Charge. Electric Potential and Potential Difference. Relation Between Electric Field Strength and Voltage. A Conductor in an Electric Field. The Electrometer. A Dielectric in an Electric Field. Ferroelectrics. The Piezoelectric Effect. Capacitance. Factors That Determine Capacitance. Capacitors. Combinations of Capacitors in Parallel and in Series. The Energy of a Charged Capacitor. Millikan's Experiment.

**18 ELECTRIC CURRENT IN METALS. DIRECT-CURRENT CIRCUITS 210**

Charge Carriers and Electric Current. Current and Current Density. The Ammeter, the Voltmeter and the Galvanometer. Closed Electric Circuit. Electromotive Force of a Power Source. External and Internal Sections of a Circuit. Ohm's Law for a Section of a Circuit Without EMF. Dependence of Resistance on Conductor's Material, Length and Cross Section. The Temperature Dependence of Resistance. Superconductivity. Equivalent Resistance. Electric Power Consumers in Series. Electric Power Consumers in Parallel. Ohm's Law for a Complete Circuit. Combinations of Cells. Ohm's Law in General Form.

**19 ELECTRIC POWER, WORK AND HEAT LOSS 232**

Electric Current and Work. Power in a Direct Current Circuit. Heating Effects of Current. Relation of Resistance to Heating Effect.

**20 THERMOELECTRICITY 237**

Thermionic Emission. Contact Potential Difference. Thermoelectromotive Force. The Peltier Effect. Application of Thermoelectricity in Science and Technology.

**21 ELECTRIC CURRENT IN ELECTROLYTES 243**

Electrolytic Dissociation. Electrolysis. Electrolysis Involving Anode Dissolution. Faraday's First Law. Faraday's Second Law. Some Applications of Electrolysis.

**22 GALVANIC CELLS AND STORAGE BATTERIES 250**

Transformation of Chemical Energy Into Electric Energy. Galvanic Cells. Polarization of Galvanic Cells and Its Reduction. Storage Batteries. Galvanic Cells and Storage Batteries in Modern Life.

**23 ELECTRIC CURRENT IN GASES AND IN VACUUM 255**

Ionization of a Gas. Dependence of Current on Voltage. Electric Discharge Through Gases at Atmospheric Pressure. Electric Discharge Through Gases at Low Pressure. Radiation and Absorption of Energy by an Atom. Cathode Rays. Plasma. Electric Current in Vacuum. The Diode. The Triode. The Cathode-Ray Tube.

**24 ELECTRIC CURRENT IN SEMICONDUCTORS 271**

Conductors, Dielectrics and Semiconductors. Pure (Intrinsic) Semiconductors. Impurity (Extrinsic) Semiconductors. P-N Junction. The Semiconductor Diode. The Transistor.

**25 ELECTROMAGNETISM 282**

Interaction of Currents. Magnetic Field as a Special Form of Matter. Magnets. Magnetic Lines of Force. Magnetic Fields in Some Simple Cases. Comparing Magnetic Properties of a Solenoid and a Permanent Magnet. Interaction Between Parallel Currents. The Permeability of a Medium. Definition of the Ampere. A Measure of the Strength of the Magnetic Field. The Homogeneous Magnetic Field. Magnetic Moment of a Current Loop. Work Done in Moving a Current-Carrying Conductor in a Magnetic Field. Magnetic Induction Due to Currents in Con-



ductors of Different Shape. Magnetic Field Strength. Paramagnetic, Diamagnetic and Ferromagnetic Substances. Magnetization of Ferromagnetic Substances. Construction of an Ammeter and a Voltmeter. The Lorentz Force Equation. Constant and Variable Magnetic Fields. Magnetic Fields in Solar and Cosmic Phenomena.

## **26 ELECTROMAGNETIC INDUCTION 310**

Flux Linkage and Inductance. Discovery of Induced Current. Induced EMF in a Straight Conductor Moving in a Magnetic Field. Faraday's Induction Experiments. Lenz's Law. The Magnitude of Induced EMF. Solenoidal Electric Field and Its Relation to Magnetic Field. Eddy Currents. Self-Induction and Self-Induced EMF. The Energy of a Magnetic Field.

### **part three**

## **Oscillations and Waves**

### **27 MECHANICAL OSCILLATIONS AND WAVES 324**

Oscillatory Motion. Conditions for Appearance of Oscillations. Classification of Oscillatory Motion Based on the Forces Acting on the Source. The Parameters of Oscillatory Motion. Quantities Characteristic of the Instantaneous State of an Oscillating Particle. Harmonic Oscillations. The Equation for Harmonic Oscillations and Its Graph. The Simple Pendulum. Laws Governing the Oscillations of a Simple Pendulum. The Compound Pendulum. Practical Uses of Pendulums. Elastic Oscillations. Energy Transformation in Oscillatory Motion. Propagation of Oscillatory Motion in an Elastic Medium. Energy Transport by Means of a Travelling Wave. Transverse and Longitudinal Waves. Waves and Rays. Wavelength. Velocity of Wave Propagation. Combination of Two Vibrations in Same Line. Reflection of Waves. Standing Waves. Interference of Waves. Mechanical Resonance.

### **28 SOUND WAVES AND ULTRASONIC WAVES 353**

What Is Sound? The Velocity of Sound. Loudness and Intensity of Sound. Pitch and Timbre of Sound. Interference of Sound Waves. Beats. Reflection and Absorption of Sound. Acoustic Resonance. Ultrasound and Its Applications.

### **29 ALTERNATING-CURRENT CIRCUITS 362**

Rotation of a Coil in a Homogeneous Magnetic Field. The Induction Generator. Effective Values of EMF, Voltage and Cur-

rent. Inductance and Capacitance in an AC Circuit. The Transformer. Induction Coil. Production, Transport and Distribution of Electric Energy.

### **30 ELECTRICAL OSCILLATIONS AND ELECTROMAGNETIC WAVES 374**

Transformation of Energy in a Closed Oscillatory Circuit. The Electron Tube Oscillator. High-Frequency Currents. Electromagnetic Field as a Special Form of Matter. Open Oscillatory Circuit. Electromagnetic Waves. Electrical Resonance. The Invention of Radio. Radiotelegraphy. Amplitude Modulation. Radiotelephony. A Simple Vacuum Tube Receiver. Radar. The Cathode-Ray Oscilloscope.

## **Optics and Special Relativity**

**part four**

### **31 THE NATURE OF LIGHT. PROPAGATION OF LIGHT 394**

Historical Survey. The Electromagnetic Theory of Light. The Quantum Theory of Light. Sources of Light. Huygens' Principle. The Velocity of Light in a Vacuum. The Velocity of Light in a Medium.

### **32 REFLECTION AND REFRACTION OF LIGHT 401**

Optical Phenomena at the Boundary Surface Between Two Media. The Laws of Light Reflection. Diffuse and Regular Reflection. The Plane Mirror. The Laws of Light Refraction. Absolute and Relative Refractive Indices. Total Reflection. Refraction by a Plane Parallel Plate and a Prism.

### **33 IMAGE FORMATION BY SPHERICAL LENSES AND MIRRORS 414**

Lenses. Focal Points and Planes. Lens Power. Image Formation for a Luminous Point Lying on the Principal Axis of a Lens. The Lens Formula. Image Formation for a Luminous Point Lying on a Secondary Axis of a Lens. Image Formation by Spherical Lenses. Lateral Magnification. Spherical Mirrors. Image Formation by Spherical Mirrors.

**34 THE EYE AND VISION. OPTICAL INSTRUMENTS 428**

Optical Systems. Deficiencies of Optical Systems, Projection Lantern. The Photographic Camera. The Eye as an Optical System. Persistence of Vision. Angle of View. Defects of Vision. Optical Illusions. The Magnifying Glass. The Microscope, Telescopes. Galileo's Telescope and Binoculars.

**35 PHENOMENA ARISING FROM WAVE NATURE OF LIGHT 447**

Interference of Light. Colours of Thin Films. Interference in a Wedge-Shaped Film. Newton's Rings. Interference in Nature and Technology. Diffraction of Light. The Diffraction Grating. Measurement of Wavelength. Polarization of Waves. Polarization of Light. Polarization of Light by Reflection and Refraction.

**36 PHOTOMETRY 464**

Energy Flux of Radiation. Solid Angle. Luminous Flux. Luminous Intensity. Illuminance. Luminance. The Laws of Lumination. Light Measurements.

**37 RADIATION AND SPECTRA. X RAYS 475**

Dispersion of Light. Dispersion by a Prism. Combining Colours. Complementary Colours. The Colour of Objects. Ultraviolet and Infrared Spectra. Ultraviolet and Infrared Radiation in Nature and Technology. Spectroscope and Spectrograph. Types of Spectra. Absorption of Light in Gases and Vapours. Kirchhoff's Law of Radiation. The Stefan-Boltzmann, Wien and Planck Radiation Laws. Solar and Stellar Spectra. Spectroscopic Analysis. The Doppler Effect. X Rays and Their Practical Uses. The Electromagnetic Spectrum. Types of Cosmic Radiation.

**38 PHENOMENA ARISING FROM QUANTUM NATURE OF LIGHT 504**

Wave and Quantum Properties of Radiation. The Pressure of Light. The Thermal Effect of Radiation. The Chemical Effect

of Radiation. Photography. External Photoelectric Effect. The Laws of External Photoelectric Effect. Einstein's Photoelectric Equation. Photocells Utilizing the External Photoelectric Effect. Internal Photoelectric Effect. Photoresistors. Photocells Utilizing the Internal Photoelectric Effect. Photocells in Science and Technology. Television. Bohr's Atom Model. The Quantized Atom. Luminescence. Lasers and Masers.

### **39 THE FUNDAMENTALS OF SPECIAL RELATIVITY THEORY 536**

Relativity in Classical Mechanics. Galilean Transformations. Experimental Foundations of Einstein's Special Theory of Relativity. What Are Simultaneous Events in Special Relativity? Lorentz Transformations. Length and Time Interval in Special Relativity. The Relativistic Velocity-Composition Law. Mass and Energy in Special Relativity. Einstein's Mass-Energy Formula. Relation Between Momentum and Energy in Special Relativity.

## **Nuclear Physics**

## **part five**

### **40 THE ATOMIC NUCLEUS 564**

Methods of Particle Detection. Radioactivity. Transmutation of Elements. Energy and Penetrating Power of Radioactive Radiation. Cherenkov Radiation. Man-Made Transmutations. The Neutron. Nuclear Structure. Nuclear Symbols and Reactions. Isotopes. Nuclear Forces. Nuclear Binding.

### **41 COSMIC RAYS. ELEMENTARY PARTICLES 585**

Cosmic Rays. The Positron. The Neutrino. The Discovery of New Elementary Particles. Classification of the Elementary Particles. Antiparticles. Mutual Transformation of Substance and Field. The Quark Model.

### **42 NUCLEAR POWER AND ITS UTILIZATION 600**

Transuranium Elements. Fission. Chain Reactions. Nuclear Reactors. Production of Power by Nuclear Reactors. Fusion. Controlled Thermonuclear Reaction. Some Applications of Radioisotopes.

**part six****Astronomy: a Brief Survey****43 THE STRUCTURE AND EVOLUTION OF  
THE UNIVERSE 618**

The Universe. The Origin and Evolution of Celestial Bodies.  
Cosmology.

**APPENDIX 632****NAME INDEX 635****SUBJECT INDEX 636**

# FOREWORD

This physics course is intended for students of technical junior colleges and gives an adequate coverage of physics at the high-school level. The aim is to provide a survey of those basics that are essential for the specialized courses that a future technician takes at college. The Soviet programme in physics for technical colleges does not include mechanics because this section of physics is studied in secondary school. But since there are many courses in mechanics, brief and extended, the teacher can always select a book that is best suited for his or her purposes. One book that we find especially useful is *Theoretical Mechanics* by E. M. Nikitin (Mir Publishers, Moscow, 1980).

The physics course that follows starts with a brief introduction about physical quantities and their measurement, the International System of Units, and the approximations that any scientist makes when measuring or calculating a quantity. It then goes on to the subject of heat and molecular physics. The other parts deal with electricity and magnetism, oscillations and waves, optics and special relativity, and nuclear physics and, finally, there is a brief survey of astronomical facts. The International System of Units is used throughout the book. However, since other systems of units are used in physics, the author has found it expedient to provide basic information about these, especially in electricity (Sections 16-9 and 16-10). To this end

the book includes an appendix whose first section is devoted to the base and derived units of the SI system.

The author, Leonid Zhdanov, wrote all the parts of the book except Part 6, which was written by Evghenii Traut. The author's son, Grigorii Zhdanov, participated in the preparation of the book for press.

Moscow, August 1979

*G. L. Zhdanov*



# Introduction

# Physical Quantities and Their Measurement

## 1-1 What is Physics!

From times immemorial man carried out systematic observations of natural phenomena, trying to record the sequence of events in nature. He learned how to forecast the course of many natural processes, for instance, the change of seasons and the swelling of rivers. This knowledge he used to choose the time to plant the seeds, to reap the harvest, and so on. Gradually people came to the conclusion that the study of natural phenomena brings them invaluable benefits.

This was the time for the appearance of scientists, that is, people who dedicated their lives to the study of natural phenomena and the generalization of the experience of former generations. They recorded the results of observations and of experiments and passed on their knowledge to their disciples. The first scientists were priests, who used their knowledge to keep the masses in subservience. Accordingly, scientists often made their recordings in a cryptographic form and carefully selected their disciples who were bound to keep the knowledge in secret.

The first books on science written for the public probably appeared in Ancient Greece. This facilitated rapid progress of science and the appearance of numerous distinguished scientists.

The word *physics* comes from the Greek word meaning nature. Therefore the science about nature was termed physics. The seventeenth century saw the beginning of a rapid progress in physics. Gradually new sciences about nature, for instance, chemistry, grew out of it. The sciences studying natural phenomena were termed natural sciences.

Long lasting studies of natural phenomena convinced the scientists of the idea of the material nature of the world around us. In the definition of V. I. Lenin (1870-1924), matter is an objective reality existing outside our mind and given to us in sensation. Hence, everything that exists in nature (and not just in our mind) is material. Thus a materialistic conception of the world is the background of our ideas about nature.

Matter exists not only in the form of material substance. For instance, radiowaves and light cannot be called substances. They represent a special form of matter, the electromagnetic field.

The study of the world around us have shown that matter is in a state of incessant motion. Any change taking place in nature is the result of motion of matter. Age-long experience convinced scientists that matter may experience transformations, but can never be created anew or annihilated. The motion of matter may also experience transformations, but motion itself is neither created nor annihilated. In other words, the world around us is matter in an incessant process of motion and development. The all-embracing measure of motion of all forms of matter is *energy*, and the permanence of motion of matter is expressed by the energy conservation law.

The most general forms of motion of matter are termed *physical*. They include: mechanical, thermal, electromagnetic, intratomic and intranuclear forms of motion of matter. Modern physics studies various forms of motion of matter, their mutual transformations, as well as the properties of substances and the field.

## 1-2 Physics and Technology

Rapid progress in the study of nature and the discovery of new natural phenomena and laws facilitated the progress in productive forces of the human society. From the eighteenth century onwards the progress in physics is accompanied by rapid progress in technology. This interrelation of the progress in physics and technology can be traced throughout modern history.

The second half of the eighteenth and the first half of the nineteenth centuries saw the appearance and perfection of the steam engine. This was accompanied by detailed study of thermal processes with the result that a new science, thermodynamics, grew out of physics. Wide use of heat engines in industry and in transportation was the reason to call this age the age of steam.

The end of the nineteenth and the beginning of the twentieth century saw the appearance and the perfection of electric machines. Simultaneously numerous new discoveries in the field of electricity were made, and physics gave birth to electrical engineering, to radio engineering and to other sciences. Wide use of electric energy in industry was the reason to term this age the age of electricity.

The period from the second decade of the twentieth century to the present day is the period of intensive research in the properties of atoms and of atomic nuclei. In this period people learned to produce nuclear energy and to use it widely in industry. The first nuclear electric power station was built in the USSR and began operating in 1954. It is now a long time since submarines and ships utilizing nuclear energy began sailing on the seas and many new atomic power stations are being built all over the world. For this reason the age in which we live may be termed the age of the atom.

The present day is the day man masters space. The world's first artificial Earth satellite was *Sputnik 1*, put into orbit in the USSR in 1957. Already in 1969 American astronauts paid a visit to the Moon. Interplanetary spaceprobes are investigating planets nearer to the Earth. Thus the end of the twentieth century marks the beginning of the space age.

Analyzing the history of natural sciences, we are bound to come to the conclusion that it is primarily physics that promotes the progress in technology and the birth of new technological fields. The advances in modern physics are the cornerstones of technological progress.

### **1-3 A Quantity and Its Measurement. Physical Quantities**

The progress in natural sciences, specifically in physics, takes the following form. A great volume of data concerning a definite group of natural phenomena is accumulated through observation and experiment. This data serves as a basis for a workable *hypothesis* (a scientific conjecture) capable of explaining the course of those events from a single viewpoint. The correctness of the hypothesis is subjected to proof

by carrying out new experiments or observations. If the hypothesis proves to be correct, it serves as a basis for a *theory*, which is called upon to explain the phenomena observed not only from a qualitative but also from a quantitative point of view and to predict new phenomena.

This means that the values calculated with the aid of the formulae derived in the theory must coincide with the results of measurements of the same quantities obtained from experiment. Therefore, experiments practically always have to include measurements of some quantities.

Everything that can be quantitatively expressed is termed a *quantity*. Thus, the length of a wire, the number of nails in a box, the speed of a boat, the temperature of water in a glass all are examples of quantities of various types. One cannot, however, compare quantities of different types. Indeed, it is impossible to answer the question as to which is greater—the length of wire or the speed of a boat. On the other hand, one can compare the length of a wire with that of a table. Should we establish that the length of wire is five times that of the table, we would adopt the length of the table as a unit of measurement since it was used for comparison with the length of the wire. The comparison of some quantity with a standard quantity is termed *measurement*. To make the result of the measurement of some quantity comprehensible to everyone this value should be compared with a standard unit of measurement (for instance, the length of an object is compared with a metre). The value of the quantity which is used for comparison with all other values of this quantity is termed *unit of measurement*. Thus, the metre is the universally accepted unit of measuring length.

Each quantity should have its own unit of measurement. The number of units contained in the measured quantity is termed the *numerical value* of this quantity. The result of measurement is expressed in the form of a concrete number, that is, the name of the unit of measurement is also stated in addition to the numeral. For instance, should the length of a piece of cloth be measured to be 5.2 metres, the 5.2 would be the numerical value of the length of this piece.

Quantities which characterize different physical properties of matter or the peculiarities of physical phenomena in nature are termed *physical quantities*.

Length, time, speed, acceleration, etc. may serve as examples of physical quantities. The numerical values of the physical quantities should always be written together with the name of the unit of measurement, for instance: 2.4 metres, 4.5 seconds, or in an abbreviated form: 2.4 m, 4.5 s.

### 1-4 Direct and Indirect Measurements

Initially each country had its own system of units of measurement, but then in the eighteenth century on recommendation of French scientists the metric system of units was devised. This system is now universally used throughout the world.

When the metric system was being created, the following units of measurement were established: the unit of length, the metre; the unit of mass, the kilogram; and the unit of time, the second. At the same time the metre and kilogram standards were produced from platinum-iridium alloy. A sample of a unit of measurement that in accordance with an international convention is regarded as the only genuine sample is termed the *standard* of this unit. The metre and kilogram standards are kept by the International Bureau of Standards in Sèvres, France. Individual countries have obtained copies of the international standards, which serve as standards of the appropriate units of measurement in the respective countries.

When the standard metre was being produced it was the intention that it should be  $1/10,000,000$ th of a quarter of the meridian passing through Paris. When the standard kilogram was being produced its mass was supposed to be equal to  $1 \text{ dm}^3$  of pure water at  $4^\circ \text{C}$ . However, more accurate measurements have shown the original metre and kilogram standards to deviate somewhat from the intended values. Because of constantly improving methods of measurement the old definitions of the metre and kilogram had to be dropped to avoid changing the standards after each new measurement. At present a new definition for the metre is adopted (see Section 35-4).

Let us find how numerical values are obtained from measurements. One can measure the length of a piece of cloth by laying a yard-stick on it, as is done in shops. Figure 1.1 shows millimetre graph paper with a rectangle drawn on it with the edges  $l = 12 \text{ mm}$  and  $b = 10 \text{ mm}$ . The area of the rectangle may be measured by placing a unit of measuring area, for instance  $1 \text{ mm}^2$ , inside it. When we count the number of squares with a 1-mm edge in the rectangle, we obtain 120, that is, the area of the rectangle is  $120 \text{ mm}^2$ .

A measurement which involves direct comparison of the value of a quantity with the unit of measurement is termed a *direct measurement*. The examples cited above are those of direct measurement of length and of area.

However, direct measurement is not always accurate. Moreover, it is not always possible or convenient. Figure 1.1

shows a circle with a diameter of 7 mm. To find the length of the circumference,  $l$ , it is more convenient to measure not the circumference itself but its diameter  $d$  and then calculate  $l$  using the formula  $l = \pi d$ .

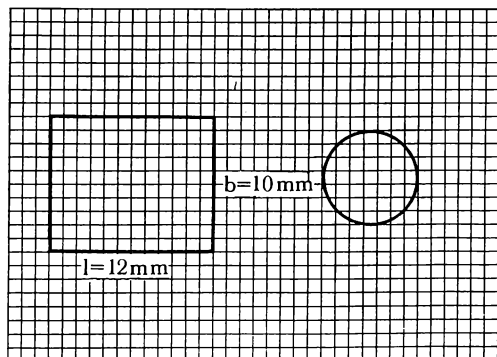


Fig. 1.1 Direct area measurement of plane figure.

When one has to measure the area of a circle he will not find it convenient to count the number of square millimetres inside the circumference. It will be easier and more accurate if he measures the diameter and calculates the area using the formula  $A = d^2/4$ , or  $A = \pi r^2$ . If one measures the length and the width of a rectangle he can compute its area using the formula  $A = lb$ .

Measurements in which a formula is used to compute the numerical value of a quantity are termed *indirect measurements*. In practice (in science and in industry) indirect measures are predominant.

### 1-5 Measurement of Angles in Astronomy

Most objects studied in astronomy are not accessible and for this reason all information concerning them has to be obtained from elaborate studies of the light (or other radiation) emanating from them. Qualitative and quantitative analysis of light will be discussed later. The important point now is that the direction of the ray of light from a heavenly body (a luminary) makes possible the determination of its position in the sky. This is done with the aid of angular measurements.

The angle which a telescope directed at a celestial body makes with the plane of the horizon is termed the *elevation*. The angle between the directions of two stars is termed the *angular distance* between them. Of course, angular distance between celestial bodies indicates only their respective posi-



tion in the sky. For instance, if the angular distance between two stars is small it should in no case be taken to mean that they are actually close to one another. One of them may be much more distant from the Earth than the other.

To make orientation by stars easier, already in ancient times people conventionally combined the brighter stars into groups—*constellations*. Later the term constellation was taken to imply parts of the sky. Astronomers take photographs of the sky and, measuring the distances between stars on the photographs, compile star catalogues, maps, diagrams and lists of accurate coordinates of the stars.

Astronomical angular measurements are made not only in the course of various astronomical observations but were used from ancient times in navigation for the purpose of orientation by the Sun and stars. Nowadays orientation by Sun and stars is used for satellites and spacecraft.

Angular measurements are also needed to determine the dimensions of celestial bodies. It is easily seen that the angular dimensions of a luminary depend on its distance. For instance, the angular diameter of the Sun, that is, the angle between the directions of diametrically opposed points of the solar disk is 0.5 degrees. The Moon is about 400 times smaller than the Sun but about the same number of times nearer to the Earth. Therefore, its angular diameter is that of the Sun, and during solar eclipses it can completely cover the solar disk. The stars, on the other hand, are so far away that even in the most powerful telescopes they are visible as bright points although it is known that some of them are much larger than the Sun.

### **1-6 Measuring Distances to Celestial Bodies by the Parallax Method**

Direct measurements of distances to celestial bodies are impossible; therefore various indirect methods have to be used. The most important among them is the *trigonometrical parallax method*.

If one looks at an object from different points (for instance, looking at the point of a pencil and shutting in turn the left and the right eye), he is bound to notice changes in its position relative to more distant background objects. The change in the apparent relative orientation of an object when viewed from different positions is termed the *parallax*. The distance between the points of observation is called the *basis* (in the example cited it is the distance between the eyes).

Having measured the parallax, we can find the distance to a faraway object. This principle is utilized in the range-finder. Here the distance between two objective lenses serves as the basis. Having determined the angle  $p$  (Fig. 1.2) between the directions of the object seen from points  $A$  and  $B$ , we can compute the distance  $D$  to the object knowing the basis  $AB = a$ . Note that from the point in which the object  $S$  is situated the basis is seen at an angle  $p$ . The distance  $D$  to the object is always much greater than the basis  $a$ , and the angle  $p$  is always very small. If the basis is perpendicular to the direction to the object, it can be taken to be equal to the arc of a circumference of radius  $D$ . Then  $a = Dp$ , where the angle  $p$  is in radians. Hence

$$D = a/p \quad (1.1)$$

Parallax measurements are used in astronomy to find the distances to celestial bodies. To measure the distance to some planet a possible method is to determine its position relative to the stars simultaneously from two observatories; the distance between these observatories will serve as the basis. However, practically it is far more convenient to make observations from one observatory at different times of the day making use of the motion of the observatory in the course of the Earth's rotation about its axis. For the sake of definiteness it was agreed to reduce parallaxes thus measured to a basis equal to the Earth's radius.

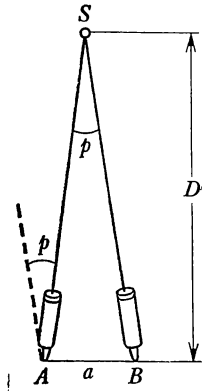
In determining distances to the stars the use is made of Earth's motion in its orbit, since the distances across the Earth prove to be too small to serve as basis. Usually the same region of the sky is photographed through a telescope at semiannual intervals. The displacement of the selected star relative to more distant stars is measured to determine its parallax, and then the distance to it is computed. In this case the distance between diametrically opposed points of the Earth's orbit from which observations are made serves as the basis. It was agreed to reduce the star parallaxes thus measured to the same basis equal to the semimajor axis of the Earth's orbit (we recall that the Earth moves along an ellipse). The parallax determined in this way is termed the *annual parallax* of a star. It is equal to the angle at which the semimajor axis of the Earth's orbit perpendicular to the star's direction is seen from the star.

If angle  $p$  is expressed in seconds of arc, we obtain

$$D = 206\,265a/p \quad (1.2)$$

since  $1 \text{ rad} = 206\,265 \text{ seconds of arc}$ . Substituting into (1.2) the value of  $a$ , we find that a distance  $D = 3.08 \times 10^{16} \text{ m}$

Fig. 1.2 Diagram of range-finder.



corresponds to an annual parallax of one second of arc. This distance is used in astronomy as a unit of length and the term for it is *parsec*\*:

$$1 \text{ parsec} = 3.08 \times 10^{16} \text{ m}$$

The distance to the star in parsecs is equal to the reciprocal of its annual parallax expressed in seconds of arc:

$$D = 1/p \quad (1.3)$$

The annual parallax of the nearest star (Alpha Centauri) turned out to be equal to 0.75 seconds of arc. The distance to it in parsecs is  $D = (1/0.75) = 1.33$  parsecs.

### 1-7 Units of Time and Their Relation to Earth's Motion

One of the most important physical quantities is *time*. Life on Earth is closely related to the periodical motion of the Sun across the sky. Therefore, the course of time and the definition of its units of measurement have for ages been connected with this motion. One of such units, the *solar day*, is the time interval between two successive passages of the Sun through the highest point above the horizon (between two middays). To measure longer periods of time the *year* is used, the time of one revolution of the Earth around the Sun. To measure small intervals of time the day was split into 24 hours, the hour into 60 minutes, and the minute into 60 seconds. Hence, the *second* is 1/86 400th of the solar day.

For a long time astronomical observations were the only method of accurate measurement of time. The invention of the clock enabled man to reproduce the units of time. The perfection of the clock led to a constant increase in its accuracy; this made it possible to establish that the diurnal rotation of the Earth is not quite uniform and that there are slight variations in the duration of the day. Therefore to establish a unit of time the use was made of the solar day averaged over a year, and the year 1900 was chosen for the sake of definiteness for it was found that the duration of the year decreases by about half a second in a century. Thus, it was agreed that the second should be equal to 1/86 400th of the solar day averaged over the year 1900.

\* The word parsec is made up of the first parts of two words: parallax and second.

Such a definition of the standard of the second is inconvenient because this standard cannot be reproduced. Progress in atomic physics made it possible to establish a new standard of the second to be discussed in Section 38-16.

### 1-8 Units of Measurement from Formulae. The International System of Units

There are many quantities in physics and each of them has its own unit of measurement. An arbitrary choice of units complicates calculations since numerical coefficients that depend only on the choice of units of measurement appear in the formulae relating different physical quantities.

Hence, when one makes an arbitrary choice of units of measurement, he has to write all the physical formulae with some proportionality factors,  $k$ . For instance, the formula for Newton's Second Law must be written in the form  $F = k_1 ma$ , and the formula for work performed by the force  $F$  along a section of the path  $s$  in the form  $A = k_2 Fs$ , etc.

However, in most formulae these factors,  $k$ , can be dropped, that is, made equal to unity, by introducing arbitrary units of measurement only for some of the physical quantities accepted as *base units* and deriving the others (called *derived units*) from formulae. Thus, in mechanics one can take length, mass, and time as the base quantities and choose units of measurement for them (for instance, the metre, the kilogram, and the second), deriving the units of measurement of other mechanical quantities from formulae. For example, let us derive the units of measurement for work and force.

The factor  $k_1$  in the formula for Newton's Second Law will be unity if with a mass of one unit and an acceleration of one unit the force will be one unit as well. With the units of measurement of mass and acceleration at hand one can choose the unit of measurement of force to satisfy this condition. Then the formula for Newton's Second Law can be written without  $k_1$ :

$$F = ma$$

We will now choose the required unit of force. To this end we substitute for  $m$  and  $a$  their units of measurement with the abbreviated designations and perform algebraic operations with the numbers as well as with the designations:

$$F = 1 \text{ kg} \times 1 \text{ m/s}^2 = 1 \text{ kg} \cdot \text{m/s}^2$$

Accept this result as a unit of measurement of force and call this unit a *newton*, the expression  $\text{kg} \cdot \text{m/s}^2$  being the *dimen-*

sions of the newton. The result obtained can be expressed as follows: 1 newton (N) is a force that imparts to a mass of 1 kg an acceleration of  $1 \text{ m/s}^2$ . Hence

$$1 \text{ N} = 1 \text{ kg} \cdot \text{m/s}^2$$

In the same way the unit for work is

$$W = 1 \text{ N} \times 1 \text{ m} = 1 \text{ N} \cdot \text{m} = 1 \text{ kg} \cdot \text{m}^2/\text{s}^2 = 1 \text{ J (joule)}$$

since  $W = Fs$ .

If the quantity whose unit of measurement is being sought is not expressed explicitly, we should find this quantity in letter form regarding the formula as an equation and then proceed by substituting known units of measurement. For instance, let it be required to derive a unit of measurement of speed from the formula  $s = vt$ . We write:

$$v = 1 \text{ m/1 s} = 1 \text{ m/s}$$

since  $v = s/t$ .

Let us now formulate a rule for deriving units of measurement. To derive a new unit of measurement of some physical quantity we must:

- (1) find a formula containing this quantity in which the units of measurement of the remaining quantities are all known;
- (2) using algebraic methods find from this formula the expression for this quantity in letter form;
- (3) substitute all known units of measurement together with their dimensions into the expression obtained;
- (4) perform all the required algebraic operations with the numbers as well as with the dimensions;
- (5) accept the result obtained as the unit of measurement sought and choose a name for it.

By the way of an example, let us derive now a unit of power.

- (1) Choose an appropriate formula:

$$W = Pt$$

- (2) find  $P$  from this formula:

$$P = W/t$$

- (3) substitute units of work and time:

$$P = 1 \text{ J/1 s} = 1 \text{ kg} \cdot (\text{m}^2/\text{s}^2)/1 \text{ s}$$

- (4) perform the algebraic operations:

$$P = 1 \text{ kg} \cdot \text{m}^2/\text{s}^3$$

(5) accept this result as a unit of measurement of power and call it the *watt* (W); the dimensions of the watt are

$$1 \text{ W} = 1 \text{ kg} \cdot \text{m}^2/\text{s}^3$$

The totality of base units together with derived units is termed a *system of units*.

It was established that to obtain a system of mechanical units it is expedient to introduce three fundamental units and to derive the rest from formulae. In the examples cited above the fundamental units were: the unit of length (metre), the unit of mass (kilogram), and the unit of time (second). The abbreviated forms m, kg, and s are the dimensions of the base units. The result of operating with these dimensions that leads to a derived unit is called the dimensions of the derived unit.

Obviously, by varying the base units of measurement (for the same physical quantities accepted as base) or by choosing other physical quantities to be the base units one can obtain numerous systems of units. Since in physics formulae are written without the factors  $k$ , calculations will yield correct results only if all the quantities are expressed in the same system of units.

Nowadays all calculations should be made mainly in the International System of Units, abbreviated SI from the French "Le Système International d'Unites". This is a unified universal system connecting units of measurement of mechanical, thermal, electric, etc. quantities. This system is built around *seven* base units:

- (i) unit of length, the metre (m);
- (ii) unit of mass, the kilogram (kg);
- (iii) unit of time, the second (s);
- (iv) unit of temperature, the kelvin (K);
- (v) unit of current, the ampere (A);
- (vi) unit of luminous intensity, the candela (cd);
- (vii) unit of amount of substance, the mole (mol).

The precise definitions of these units will be presented later, and the units of speed, force, work, and power in the SI system were presented above. The units of measurements in this system are presented in Appendix.

## 1-9 Treatment of Data

Usually measurements are made with the aid of instruments. For instance, measurements of length can be made with the aid of a ruler with divisions on it, a slide caliper, a micrometer screw gauge, and other instruments. All the

results obtained in the measuring process are approximate. The errors of measurement are due to the deficiencies of the instruments themselves, depending, in addition, on the individual qualities of the operator, for instance on his attentiveness, eyesight, etc. Only numbers obtained in counting objects can be accurate, but even so there is the additional condition that they remain finite and constant during the time the counting is made.

Errors of measurement due to the inaccuracy of the instrument can easily be taken into account since the maximum error determined by the instrument's design is stated on the instrument. If it is not, the error is usually taken to be half of the smallest division on the instrument's scale. For instance, the error in measuring length with a millimetre ruler may be assumed to be equal to 0.5 mm; when a slide caliper is used, the error may be assumed to be equal to 0.05 mm if its vernier has 10 divisions to a millimetre.

*Random errors* are due to incorrect position of the eye, to carelessness of the operator, etc. and can be substantially diminished by repeating the measurements and accepting the *arithmetic mean* of all the measured values as the *true* value of the measured quantity.

Suppose the values obtained in several measurements were  $a_1, a_2, a_3, a_4, a_5$ ; in that case the true value of this quantity is taken to be

$$a_{\text{mean}} = \frac{a_1 + a_2 + a_3 + a_4 + a_5}{5}$$

For the general case of  $n$  measurements this is written in the form

$$a_{\text{mean}} = \frac{1}{n} \sum_{i=1}^n a_i \quad (1.4)$$

where

$$\sum_{i=1}^n a_i = a_1 + a_2 + \dots + a_n$$

The magnitude of the difference between the true value of the quantity being measured and the result of an individual measurement is termed *absolute error* of this measurement and is denoted  $\Delta a$ . Let us agree to use the arithmetic mean of the results of measurements of a quantity instead of its unknown true value. Then the values of the absolute errors in our example will be

$$\Delta a_1 = a_{\text{mean}} - a_1, \quad \Delta a_2 = a_{\text{mean}} - a_2, \quad \text{etc.}$$

The arithmetic mean of all the absolute errors is accepted as the final absolute error of the value of the quantity  $a$ , that is,

$$\Delta a_{\text{mean}} = \frac{|\Delta a_1| + |\Delta a_2| + |\Delta a_3| + |\Delta a_4| + |\Delta a_5|}{5}$$

In general

$$\Delta a_{\text{mean}} = \frac{1}{n} \sum_{i=1}^n |\Delta a_i| \quad (1.5)$$

where

$$\sum_{i=1}^n |\Delta a_i| = |\Delta a_1| + |\Delta a_2| + \dots + |\Delta a_n|$$

The result of a single measurement, with account taken of the fact that the deviation from the arithmetic mean may be of either sign, can be written in the form

$$a = a_{\text{mean}} \pm \Delta a_{\text{mean}} \quad (1.6)$$

This means that  $a$  lies in the interval

$$a_{\text{mean}} - \Delta a_{\text{mean}} \leq a \leq a_{\text{mean}} + \Delta a_{\text{mean}}$$

Note in addition that absolute error is a denominate number.

The absolute error does not present a complete picture of accuracy of measurement. In practice the accuracy of measurement of a quantity  $a$  is additionally assessed with the aid of the relative error  $\delta a$ . Relative error is the term for the ratio of the absolute error to the value of the quantity being measured in per cent:

$$\delta a = \frac{\Delta a}{a} \times 100\% \quad (1.7)$$

In repeated measurements arithmetic means of  $a$  and  $\Delta a$  are substituted into formula (1.7). In cases when no great accuracy is required the acceptable relative error in engineering is under five per cent. Such, for instance, is the accuracy of the slide rule, which the student is recommended to use for solution of problems.

Let us now find how approximate numbers are recorded with the aid of significant digits. Suppose in measuring length we obtained the result  $l = 12.31 \text{ cm} \pm 0.18 \text{ cm}$ . There is an error already in the tenths of a centimetre; therefore there is no sense in writing out the hundredths. A more correct way is to write  $l = 12.3 \text{ cm} \pm 0.2 \text{ cm}$ . Note that the last numeral 3 is unreliable since it may be anything from 1 to 5. Since in this digit the error is less



than one half of a centimetre, the number of integral centimetres is conventionally accepted as the last significant digit of this number, the order of the numerals being from left to right.

Thus, in the above example there are two significant digits. The last significant digit of an approximate number is the right-hand digit for which the absolute error remains less than one half of the digit, the error exceeding one half of the following digit.

Writing the approximate number 12.3 cm in millimetres (123 mm) or in metres (0.123 m) does not change the number of significant digits; therefore, the zeros preceding the decimal point do not represent significant digits.

Quite frequently the absolute error is, for the sake of brevity, not recorded with the approximate number; only the significant digits are recorded. In our example we should write  $l = 12$  cm. This means that the error of the number does not exceed 0.5 cm.

Since the relative accuracy of an approximate number is independent of the position of the decimal point, it is expedient to place it after the first digit, multiplying the number by the appropriate power of 10 to retain its magnitude. For instance, it is better to write  $1.2 \times 10$  cm instead of 12 cm. Here are additional examples: for  $m = 0.0543$  kg we write  $m = 5.43 \times 10^{-2}$  kg, for  $t = 324.2$  s we write  $t = 3.242 \times 10^2$  s, etc. Such a way of writing makes the number of significant digits obvious at first glance, and when operating with approximate numbers one may easily check the order of magnitude of the number obtained.

Note, in addition, that there are three significant digits in, say,  $1.20 \times 10^2$  J and that the zero after the decimal point shows that hundredths of a joule were measured but were not registered. Such zeros belong to significant digits and should not be omitted.

## 1-10 Combining Errors

When performing algebraic operations with approximate numbers one should know how to find the error in the result of calculations. Suppose we know the errors of the numerical values we substitute into a formula. In that case we obtain as a result of computation an approximate number whose error may be determined with the aid of several theorems. We present these theorems without proof.

*Theorem 1.* The absolute error of the algebraic sum of approximate numbers is the arithmetic sum of the absolute

errors of all the addends. Let  $x = a + b - c$ . Then

$$\Delta x = \Delta a + \Delta b + \Delta c \quad (1.8)$$

*Theorem 2.* The relative error of a product of approximate numbers is the arithmetic sum of the relative errors of all the multipliers. Let  $x = a \times b \times c$ . Then

$$\frac{\Delta x}{x} = \frac{\Delta a}{a} + \frac{\Delta b}{b} + \frac{\Delta c}{c} \quad (1.9)$$

*Theorem 3.* The relative error of a fraction is the arithmetic sum of the relative errors of the numerator and of the denominator. Let  $x = a/b$ . Then

$$\frac{\Delta x}{x} = \frac{\Delta a}{a} + \frac{\Delta b}{b} \quad (1.10)$$

*Theorem 4.* The relative error of a power is the absolute value of the power multiplied by the relative error of the base. Let  $x = a^{-n/m}$ . Then

$$\frac{\Delta x}{x} = \frac{n}{m} \frac{\Delta a}{a} \quad (1.11)$$

Here is an example. Let  $x = \frac{a^2 \sqrt[3]{b}}{a^4 \sqrt{d^3}}$ . Applying Theorems 2, 3 and 4, we obtain

$$\frac{\Delta x}{x} = 2 \frac{\Delta a}{a} + \frac{1}{3} \frac{\Delta b}{b} + 4 \frac{\Delta c}{c} + \frac{3}{2} \frac{\Delta d}{d}$$

When there is no need for an accurate appraisal of the error of computations, the following approximate rule may be used. If there is no addition or subtraction in the original formula, the number of significant digits retained in the answer should coincide with that in the approximate number with the least number of significant digits. For instance, the division of the approximate number  $5.74 \text{ m}^2$  by  $1.2 \text{ m}$  should be done to two significant digits in the quotient. Dividing,

$$5.74 \text{ m}^2 \div 1.2 \text{ m} = 4.783... \text{ m}$$

However, it suffices to write  $5.74 \text{ m}^2 \div 1.2 \text{ m} = 4.8 \text{ m}$ .

The students are advised to follow this rule when solving problems and doing laboratory work.

## 1-11 Density of Substance

Measurements of the mass and of the volume of bodies made of the same substance show their mass to be directly proportional to their volume. In mathematics the direct proportionality of the variables  $x$  and  $y$  is expressed by the formula  $y = Kx$ , where  $K$  is the proportionality factor that remains constant as  $x$  and  $y$  vary.

The proportionality factor in a formula of physics characterizes a particular property of the process. It remains constant only if definite conditions are met. One of the most important tasks in the study of the laws of nature is to reveal the physical meaning of the proportionality factors in the formulae and the determination of their values in specific conditions. We also note that the numerical values of the proportionality factors depend on the choice of units of measurement (see Section 1-8).

The dependence of the mass of a body on its volume discussed above may be expressed by the formula

$$m = KV \quad (1.12)$$

Here  $K$  depends on the choice of the units of measurement and on the substance, since the mass of the body depends not only on the volume but on the kind of substance the body is made of as well. Usually in physics the dependence of the proportionality factor on the choice of units of measurement is expressed with the aid of a separate multiplier  $k$ . Therefore,  $K$  in (1.12) can be represented by the product of two multipliers:  $k$ , expressing the dependence of mass on the choice of the units of measurement, and  $\rho$ , expressing the dependence of the mass on the substance constituting the body:

$$K = k\rho \quad (1.13)$$

Then formula (1.12) assumes the form

$$m = k\rho V \quad (1.14)$$

As was stated above,  $k$  in (1.14) may be dropped if it is used to derive a new unit of measurement. Therefore formula (1.14) is written as

$$m = \rho V \quad (1.15)$$

The quantity  $\rho$  expressing the dependence of mass on the substance and on external conditions is termed *density of substance*. The measure of density is the mass contained in a unit of volume

$$\rho = m/V \quad (1.16)$$

The volume of a body changes with pressure and temperature and this means that density depends on external conditions. Let us derive now a unit of density:

$$\rho = m/V, \quad \rho = 1 \text{ kg}/1 \text{ m}^3 = 1 \text{ kg}/\text{m}^3$$

In the SI system the unit of density is defined as the density of such a substance  $1 \text{ m}^3$  of volume of which has a mass of 1 kg. For calculations the density of substances is taken from appropriate tables.

part one

# **Heat and Molecular Physics**

# The Fundamentals of Kinetic Theory of Matter

## 2-1 First Principles of Kinetic Theory

Whenever we observe natural phenomena or study the properties of various kinds of substances we are bound to be impressed by the great variety of ways in which matter manifests itself. We may understand why the properties of matter are so different if we study its internal structure.

In the beginning of the last century the British chemist and physicist John Dalton (1766-1844) demonstrated that many laws governing natural phenomena can be explained with the aid of the concept of an *atom* and a *molecule*. He laid the scientific foundation for the theory of molecular structure of matter. Towards the beginning of this century the kinetic theory of matter was completed and proved by numerous experiments. What is the essence of this theory?

All substances consist of *molecules* (from the Latin *moles* meaning mass and *cule* a diminutive suffix). Molecule of any substance is the smallest particle of that substance which can exist and still possess the chemical properties of the bulk material.

Molecules are made up of *atoms* (from the Greek *atomos* meaning indivisible), for instance, a molecule of water is made up of two atoms of hydrogen and an atom of oxygen, and this is written in the form  $\text{H}_2\text{O}$ . A substance retains its chemical properties in the process of some natural phenom-

enon if its molecules remain unchanged. If, on the other hand, the molecules change their structure or split into separate atoms, new kinds of substances with different chemical and physical properties are produced. For example, water molecules can be decomposed into atoms of hydrogen and oxygen. Efforts to decompose these gases into still more elementary substances using chemical methods proved unsuccessful.

Substances that cannot be decomposed into more elementary substances with the aid of chemical methods are termed *chemical elements*, for example oxygen, nitrogen, and lead. Every chemical element has its proper place (and proper number) in the Mendeleev Periodic Table. Atoms combined in a group form a molecule of substance. The totality of identical molecules forms a definite kind of material. The chemical and physical properties of this material are determined by the number and the kind of atoms contained in its molecules.

The properties of a substance depend also on the mutual arrangement of the atoms inside it. For instance, both graphite and diamond are made up of carbon atoms, and the only difference in their internal structure is the mutual arrangement of atoms. Yet the physical properties of these substances are quite different: diamond is very hard, transparent to light, and an excellent insulator; graphite, on the other hand, is soft, not transparent, and a conductor. Lastly, the properties of a substance are influenced by the surroundings. All this is due to the fact that molecules and atoms constantly interact and possess chemical energy. Motion and interaction of atoms and molecules is the cause of the boundless variety of the majority of the observed natural phenomena.

Let us now formulate the principal points of the kinetic theory of matter:

(1) all substances are made up of molecules separated by intermolecular distances;

(2) the molecules of a substance are constantly in a state of random motion;

(3) at small intermolecular (interatomic) distances both attractive and repulsive forces act between them, the origin of those forces being electromagnetic.

Let us recall some phenomena in support of these statements.

Experiments show that all gases are easily compressible. This proves that there is substantial free space between the molecules of a gas. Liquids and solids are also compressible, but considerably less than gases are. This means that in

liquids and in solids there are intermolecular distances as well, but they are much smaller than in gases.

Mutual penetration of molecules of one substance of the space between the molecules of another results in the mixing of various gases, liquids, in dissolution of solids in liquids, and in the evaporation of liquids and solids.

The tendency of the gas molecules to occupy the free space shows that the gas molecules are in a state of constant random motion.

A special point to note is that random motion of molecules is often termed *thermal motion* because it is closely related to the concept of temperature, which will be discussed now.

## 2-2 Concept of Temperature

The sense of touch helped man to obtain the primary notion of temperature. Touching various bodies with our hands we sometimes say "a cold body" or "a warm body" to denote the degree of heating to which the body had been subjected. The quantity characterizing the degree of heating to which the body had been subjected is termed the *temperature* of this body.

The determination of temperature by touch is not objective and may be the cause of error. Indeed, if we press a naked hand against two objects lying on a table, of which one is made of plastic and the other of metal, we will say that the metal object is cooler although in actual fact the temperature of both objects is the same. Therefore to measure temperatures objectively a special instrument termed *thermometer* was invented. The operation of conventional medical thermometers as well as of thermometers to measure the temperature of air is based on the expansion of bodies upon heating and on their contraction upon cooling. Note that there are thermometers whose operation is based on the utilization of other properties of matter. They will be described later.

The temperature of every body is closely related to the energy of motion of its molecules. The physical meaning of the concept of temperature is as follows: the higher the temperature of the body, the greater the average kinetic energy per molecule of this body. Therefore, to heat a body energy must be supplied to it, and to cool the body energy must be taken away from it.

Experience shows that when two bodies with different temperatures are in contact, the body with the higher temperature always gets cooler and the body with the lower tem-

perature always gets hotter. This means that energy exchange takes place between the bodies, ceasing when the temperatures of the bodies become equal. Such exchange of energy is termed *heat exchange*. Hence, by measuring the temperatures of the bodies we can find out beforehand which bodies shall receive heat in the course of heat exchange and which shall give it away. If the temperatures of the bodies measured will prove to be equal, this will mean that the energy of each of them remains constant. This makes temperature a valuable concept.

Since the kinetic energy of a molecule is proportional to the square of its velocity, heating a body results in the increase in the average velocity of its molecules; the result of cooling is a decrease in it.

A person's sense of heat and cold is also due to heat exchange between him and the surrounding medium. If the temperature of the object you touch exceeds that of your hand, energy is transmitted from the object to the hand. If, on the other hand, the object's temperature is lower, heat is transmitted from your hand to the object. The more rapid this exchange of energy, the more intense is your perception of heat or cold.

Now we can understand why the metal object appeared to be colder than the plastic one. The explanation is that the energy transfer from the hand to the metal object was more rapid than in the case of the plastic object and the temperature of the hand in the first case dropped quicker than in the second. (Ponder on the question of what you feel when you touch the same objects in the case when their temperature, identical for the two objects, exceeds the temperature of your body.)

### 2-3 Diffusion

One of the wide-spread natural phenomenon which may be explained by the random motion of molecules is diffusion (from the Latin *diffusus* meaning spreading). The spreading of the odour of flowers or of food in the process of cooking may serve as examples of diffusion. The concentration of aromatic molecules near a bouquet of flowers is great because the evaporation and random thermal motion of the molecules results in the mixing of the air molecules with those of aromatic substances, which leads to the spreading of the odour throughout the room. Such mixing results in the equalization of the concentration of aromatic molecules throughout the space of the room.



The process of equalizing concentrations of some substance in space due to random motion of the molecules is termed *diffusion*. It was mentioned above that the average velocity of molecules rises with temperature. Therefore, diffusion of noninteracting molecules must become more intense with the rise in temperature, which agrees with all experiments on diffusion in gases.

Diffusion in liquids is much slower than in gases. If we pour water into a vessel and then carefully pour coloured alcohol on top of it so that a clear boundary is visible, after several days we shall observe that the boundary exists no longer and that water became coloured to a substantial depth on account of mutual diffusion of water and alcohol. (Ponder on the question why this experiment cannot be explained by the action of the force of gravity.)

Diffusion in solids at room temperatures is so slow that can usually be detected only after many months. However, if one presses two plates made of different metals tightly together and keeps them at a temperature of several hundreds degrees, interdiffusion of the metals will become noticeable already after several hours. This can be done by separating the plates and investigating their surface.

What is the reason for diffusion to be slower in liquids than in gases and still slower in solids than in liquids? The obvious explanation is that in liquids the molecules are closer to one another than in gases and the forces of molecular attraction retard the process of diffusion. Since the molecular forces in solids are stronger than in liquids, the diffusion process in solids is even slower.

## 2-4 Forces of Molecular Interaction

Studies of the structure of matter lead to the conclusion that all molecules include both positive and negative electric charges. Since like charges are repelled and unlike charges attracted, forces of attraction and of repulsion act simultaneously between the molecules. Besides, there is a magnetic interaction between the moving charges in atoms and molecules, which contributes to the resultant of the molecular forces of attraction and repulsion.

Molecular forces in solids manifest themselves in the form of elastic forces during various deformations and in the form of forces responsible for the strength of bodies. Those forces act only at very short distances. This is the reason why, for instance, it is impossible to mend a broken porcelain cup by pressing together its pieces, for in this case the distance

between the overwhelming majority of molecules across the break is so great that the molecular forces between them cannot act. However, if one takes a ductile material, he can bring a large number of molecules close enough by pressing parts of it together. In that case the adhesive forces between the parts of the object pressed together will be so great that the parts cannot easily be separated. For instance, one may join lead rods securely by pressing together their well cleaned plane surfaces. For a  $1\text{-cm}^2$  contact surface such a composite rod will hold a 5-kg weight (Fig. 2.1a).

Objects made of hard materials can be securely joined by pressing together finely polished surfaces of those objects. (Ponder on the principles of welding and of glueing solids together.) Molecular forces are one of the reasons of cohesion of gauge plates (Fig. 2.1b).

Forces of interaction act when any two molecules (of the same and different types) are brought together, and the magnitude of those forces depends on the nature of the molecules.

Figure 2.2 shows a typical diagram of the dependence of the net force  $F_{\text{net}}$ , which of course is the resultant of the

Fig. 2.1 Experiments demonstrating attraction of molecules: (a) lead rods hold 5-kg weight; (b) gauge plates.

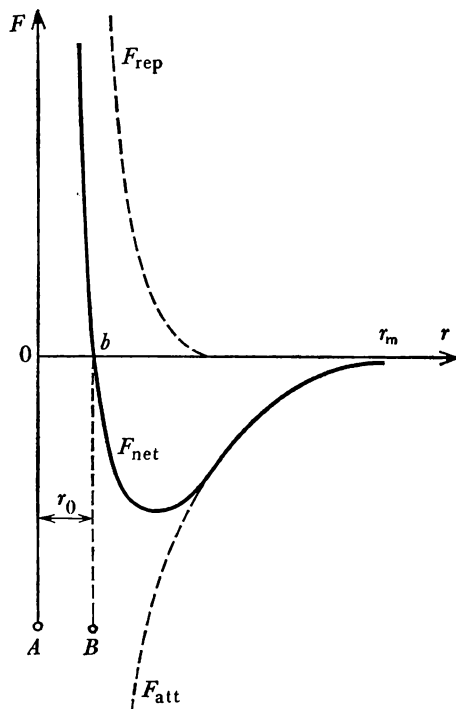
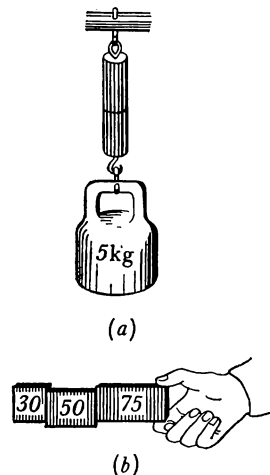


Fig. 2.2 Net molecular force acting on molecule B versus distance  $r$  between molecules A and B.

attractive and of repulsive forces acting between the molecules  $A$  and  $B$  on the distance  $r$  between the centres of the molecules. In plotting the diagram the forces of repulsion  $F_{\text{rep}}$  are assumed to be positive and the forces of attraction  $F_{\text{att}}$  negative.

The distance  $Ob = r_0$  corresponds to the state of stable equilibrium of the molecules, for at this point the resultant of the molecular forces is zero and any variation of the intermolecular distance leads to the appearance of forces striving to return the molecules to their former positions. Indeed, it may be seen from the diagram that when the intermolecular distance is made shorter than the equilibrium distance, the forces of repulsion (positive) become prevalent, and when it is made longer, the prevailing forces will be those of attraction (negative).

The minimum intermolecular distance at which forces of molecular interaction are so small that they may be neglected is termed the *radius of molecular action* ( $r_m$  in Fig. 2.2). It is of the order of one nanometer ( $1\text{nm} = 10^{-9}\text{ m}$ ), which is ten angstroms ( $1\text{ \AA} = 10^{-10}\text{ m}$ ). Note that molecules in a substance possess potential energy due to the forces of molecular interaction.

## 2-5 Kinetic and Potential Energies of Molecules

If we denote the mass of a molecule of a body by  $m$  and the velocity of its translational motion by  $v$ , the kinetic energy of its translational motion is

$$K_{\text{trans}} = \frac{mv^2}{2}$$

The molecules of a body possess different velocities and energies  $K_{\text{trans}}$ . Hence to characterize the state of the body, the average energy of translational motion  $\overline{K}_{\text{trans}}$  is used:

$$\begin{aligned}\overline{K}_{\text{trans}} &= \frac{K_{\text{trans } 1} + K_{\text{trans } 2} + \dots + K_{\text{trans } N}}{N} \\ &= \frac{1}{N} \left( \frac{m_1 v_1^2}{2} + \frac{m_2 v_2^2}{2} + \dots + \frac{m_N v_N^2}{2} \right)\end{aligned}$$

where  $N$  is the total number of molecules in the body.

If all the molecules are identical, then

$$\overline{K}_{\text{trans}} = \frac{m}{2} \left( \frac{v_1^2 + v_2^2 + \dots + v_N^2}{N} \right) = \frac{m}{2} v_{\text{rms}}^2 \quad (2.1)$$

where  $v_{\text{rms}}$  is the *root-mean-square speed* of random molecular motion:

$$v_{\text{rms}} = \sqrt{\frac{v_1^2 + v_2^2 + \dots + v_N^2}{N}} = \sqrt{\frac{1}{N} \sum_{i=1}^N v_i^2} \quad (2.2)$$

As was stated above, the molecules in a body possess potential energy in addition to the kinetic energy. Let us presume the potential energy of an isolated molecule not interacting with other molecules to be zero. In that case the potential energy of interaction of two molecules due to forces of repulsion will be positive and that due to forces of attraction negative (Fig. 2.3a), since in order to bring the molecules

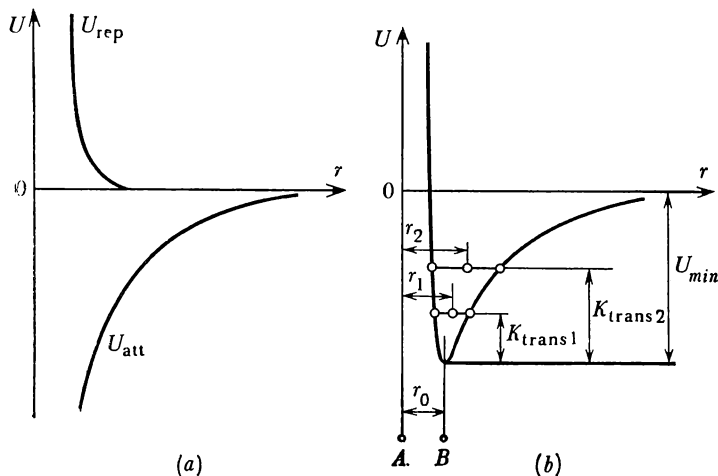


Fig. 2.3 (a) Dependence of potential energy of attraction and repulsion of two molecules on intermolecular separation; (b) resultant potential energy of interaction of two molecules.

closer work should be performed against the forces of repulsion, whereas the forces of attraction perform work themselves. Figure 2.3b shows the dependence of the potential energy,  $U$ , of interaction of two molecules on the distance  $r$  between them. The part of this diagram around the minimum of  $U$  is called a *potential well*, the minimum energy  $U_{\text{min}}$  being the *depth* of the potential well.

In the absence of kinetic energy  $K_{\text{trans}}$  the molecules would occupy positions at a distance  $r_0$  corresponding to the state of stable equilibrium, since in this case the resultant of intermolecular forces is zero and the potential energy is minimal (see Fig. 2.2). To bring the molecules apart, work should be done against the forces of molecular interaction equal in magnitude to  $U_{\text{min}}$ , in other words, the molecules will have to surmount a potential barrier  $U_{\text{min}}$  high.

Since molecules always possess some kinetic energy, the distance between them changes constantly and may turn out to be either greater or less than  $r_0$ . If the kinetic energy of molecule  $B$  will be less than  $U_{min}$ , for instance  $K_{trans}$  in Fig. 2.3b, the molecule will move inside the potential well.

Molecule  $B$  may, moving against the forces of attraction or repulsion, attain positions either close or away from  $A$  in which its kinetic energy  $K_{trans}$  will be transformed into the potential energy of molecular interaction. Those extreme positions of the molecule are determined by the points on the potential curve corresponding to the level  $K_{trans 1}$  from the bottom of the potential well (see Fig. 2.3b). Subsequently the forces of attraction or repulsion will push the molecule  $B$  away from those extreme positions. In this fashion the forces of interaction retain the molecule at some average distance  $r_1$ .

If the kinetic energy of molecule  $B$  is greater than  $U_{min}$ , the molecule will surmount the potential barrier and the distance between the molecules can increase to infinity.

When a molecule moves inside a potential well, its kinetic energy ( $K_{trans 1}$  and  $K_{trans 2}$  in Fig. 2.3b) rises with the rise in the temperature of the body, this being accompanied by the increase in intermolecular distance ( $r_1$  and  $r_2$ ). This explains the thermal expansion of solids and liquids upon heating.

The cause for the increase in the average intermolecular distance is that the potential diagram to the left of  $U_{min}$  rises much steeper than to the right. Such asymmetry is due to the fact that the forces of repulsion decrease with an increase in  $r$  much more rapidly than the forces of attraction (see Fig. 2.2).

## 2-6 Concept of Internal Energy

In the study of heat exchange a question arises as to what the constituents of the energy of a body are and how these constituents change in the course of heat exchange.

It was established above that molecules of a body possess kinetic and potential energy. Atoms inside a molecule and electrons inside an atom, too, possess kinetic and potential energy the term for which is *chemical energy*. Atomic nuclei possess enormous energy, termed *nuclear*. The sum of the kinetic and potential energies of all the particles constituting a body is termed *internal energy* of that body.

It was established that there is a constant energy exchange between individual parts of a body, but in the absence of

external influences its internal energy remains constant. Experiments have shown that the internal energy of bodies are determined solely by their state, no matter how this state is achieved. Therefore internal energy of a body is often said to be a function of state. A system of bodies whose internal energy remains constant is termed *closed system*.

Molecular physics deals only with phenomena in which molecules remain unchanged. In such phenomena all variations of the internal energy of a body are due entirely to the variations of the kinetic and potential energies of its molecules. Of practical importance is primarily the variation of the internal energy of a body. Therefore in those cases the term internal energy of a body is taken to imply only the sum of the kinetic energies of all of its molecules and of the potential energies of their interaction (in short, the sum of the molecular-kinetic and molecular-potential energies of the body).

### 2-7 Probability of an Event. The Statistical Method

Molecular physics studies phenomena resulting from the motion and interaction of so great a number of molecules that it defies imagination. For instance, a thimble contains so many molecules of air that it would take several billion years to count them. Therefore it is a hopeless task trying to measure kinetic energies of all the individual molecules. Even if we were to know the kinetic energies of all those molecules, we would not be able to use those numbers for practical purposes for there would be too many of them and it would take ages to perform arithmetical operations with them. Here even the computer would be of little help. To study phenomena of this sort use is made of theory of probability. Let us learn about some of its conclusions and find the meaning of the probability of an event.

Suppose identical objects, for instance, balls of different colour, 4 black, 6 red and 10 white, lie in a box without a lid. Let us cover the box with a handkerchief, shake the box, and take one ball out of the box without looking. What will be its colour? Since the white balls are the most numerous one could expect it to be white. However, after we take a look at the ball it may turn out to be red. Can numerical relations be established for cases of such sort? The answer is positive.

Let us term the extraction of a ball an *event*. The total number of balls in the box is 20. Since they cannot be distin-

guished by touch, the chances to extract any ball are identical. Such events of which not one has any advantage over the other are termed equally probable. Imagine the balls to be numbered from 1 to 20. Should we extract a ball number 5, it would be impossible to confuse it with any other ball contained in the box. If the realization of an event automatically precludes the realization of another, they are termed *incompatible*.

Hence, extracting blindly a ball, we realize one of the 20 equally probable incompatible events. Of this number 4 events favour the extraction of a black ball, 6 the extraction of a red ball and 10 the extraction of a white ball. The ratio of the number of events favouring the expected event to the total number of equally probable and incompatible events is termed *probability* of the expected event,  $p$ . Hence the probability of extracting a black ball is  $4/20$ , or  $p_b = 1/5 = 0.2$ . In the same way for the red and white balls we obtain  $p_r = 6/20 = 3/10 = 0.3$  and  $p_w = 10/20 = 1/2 = 0.5$ .

Let us now make the following experiment. We take out a ball, record its colour and throw it back into the box. Then we mix the balls, extract one again, record its colour and put it back again. We repeat this trial 23 more times and count the number of times we extracted black, red and white balls. Finally, we divide those numbers by the number of trials. Suppose we had obtained the following numbers for the black, red and white balls:  $A'_b = 0.24$ ,  $A'_r = 0.36$ , and  $A'_w = 0.40$ . Repeating the series of 25 trials once more, we shall obtain quite different numbers, say,  $A''_b = 0.12$ ,  $A''_r = 0.20$ , and  $A''_w = 0.68$ . However, if we repeat the series of trials 1000 times, the values of  $A_b$ ,  $A_r$ , and  $A_w$  obtained would be quite close to  $p_b$ ,  $p_r$  and  $p_w$ . Actually the difference between  $p$  and  $A$  decreases with the increase in the number of trials and for a sufficiently large number of trials may be as small as desired (in magnitude). This very important relation is proved in the theory of probability. The term for it is the *law of large numbers*: if equally probable incompatible events are repeated a great number of times, the relative number of realizations of each of them will be close to the probability of the event, the difference being the smaller the greater the number of repetitions.

Note in addition that the probability of a certain event which is sure to happen is unity. In the above example the probability to extract a black or a red or a white ball is unity, for there are no balls of other colour in the box and an extracted ball will, of necessity, be of one of those three colours. Therefore  $p = 20/20 = 1$ .

In the evolution of the kinetic theory a particular method of studying natural phenomena played an important part. Regularities that are realized in experiments only when a great number of identical independent phenomena are involved are termed statistical and the method of establishing them is the *statistical method*.

A characteristic feature of the statistical method is the determination of the *average value* for a multitude of identical phenomena, because statistical regularities hold only for such average values. Phenomena studied in molecular physics involve the motion and interaction of a great number of molecules. Therefore, only average values characteristic of the world of molecules are used to describe the molecules.

To sum up, the foundation of molecular physics is the statistical method of studying natural phenomena, which uses the theory of probability to draw its conclusions.

## Kinetic Theory of Gases

## 3

### 3-1 The Gaseous State

As we know, the properties of a substance are determined by the motion of its molecules and by the forces of interaction between them. The forces of molecular interaction strive to hold the molecules at definite distance from one another, whereas random motion makes them disperse in space, that is, makes the volume occupied by the substance greater. The shape and the volume of a body is determined by the combined action of those factors.

We recall that the most important factor determining the behaviour of a gas is the random motion of its molecules. The average kinetic energy of gas molecules  $\bar{K}_{\text{trans}}$  substantially exceeds their energy of interaction,  $U_{\text{min}}$  (see Fig. 2.3); the forces of interaction are incapable of holding the molecules together and they fly about the space allowed to the gas. In such circumstances the average distance between the molecules is determined by the number of molecules and the dimensions of the vessel containing the gas.

The space limited by a spherical surface of a radius equal to the radius of molecular interaction is termed the *sphere of molecular interaction*. Only molecules whose centres are inside this sphere interact with a given molecule.

Calculations show that the average distance between the molecules of a gas in normal conditions are about 3 nm,



the radius of molecular interaction being equal to approximately  $r_m \approx 1$  nm (see Section 2-4). Therefore, if we draw spheres of molecular interaction around all the molecules of the gas at some instant of time, the combined volume of those spheres will prove to be only a negligible fraction of the total volume occupied by the gas, and the majority of the molecules will find themselves outside the spheres of interaction of other molecules. This means that for the most part the gas molecules move independently, their motion being rectilinear (due to inertia), until they collide with other molecules (or with the walls of the vessel). Collision changes the magnitude and the direction of the molecule's velocity, and the molecule moves with a new constant velocity until it is involved in a new collision.

Forces of repulsion act between colliding molecules and the magnitude of these forces is determined by  $K_{\text{trans}}$ . The greater  $K_{\text{trans}}$ , the greater the forces of repulsion acting in the process of collision. For  $K_{\text{trans}} \gg U_{\text{min}}$  they greatly exceed the forces of attraction, which initially appear as the molecules come closer together. Therefore the forces of attraction between the molecules of a gas can usually be neglected. Acted upon by powerful forces of repulsion the molecules after collision fly away in different directions.

For  $\overline{K}_{\text{trans}} \gg U_{\text{min}}$  the interaction and random motion of gas molecules are influenced little by the shape of the potential well and by the magnitudes of  $U_{\text{min}}$  and  $r_0$ , which depend on the nature of the gas. This explains why the properties of different gases (in normal conditions) are similar.

However, when the gas is compressed so that the average distance between its molecules approaches the radius of molecular interaction  $r_m$  the forces of molecular attraction can no longer be ignored. Also, when the gas is cooled to such low temperatures that the condition  $\overline{K}_{\text{trans}} \gg U_{\text{min}}$  no longer holds, the forces of molecular attraction become important. In both cases various gases exhibit peculiar properties.

To sum up, the peculiarity of the gaseous state of matter (at pressures not too high and at temperatures not too low) is that at any instant of time the number of interacting gas molecules is a negligible fraction of the total and that in the interaction itself mutual attraction of the molecules can be ignored.

One should not imagine the thermal motion of the gas molecules to be solely translational. Additional rotational motion of a gas molecule made up of several atoms may be excited as a result of collisions. The velocity of rotational

motion increases with temperature in the same way as that of translational, which means that it, too, is thermal motion. Thus, thermal motion of multiatomic gas molecules consists of translational and rotational motion.

Note that the atoms in a molecule may additionally take part in vibrational motion, but at low and medium temperatures it is insignificant. Only at very high temperatures does the contribution of vibrational motion of atoms in gas molecules become noticeable.

### 3-2 Brownian Motion

One of the results of random motion of gas molecules is the Brownian motion of solid particles suspended in a gas or in a liquid.

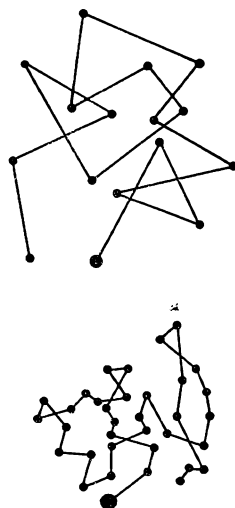
Imagine a particle so small that it can be seen only in a microscope. Let it be among gas molecules moving at random. Such a particle is a giant compared to the molecules, and the moving molecules will strike against it from all sides and in different directions. Such impacts will create a pressure on the surface of the particle similar to the pressure that rain drops create on the surface of an open umbrella.

It follows from the theory of probability that the greater the particle in comparison with a molecule the closer to zero is the resultant of the impact forces of the molecules striking against it at any instant of time. However, because of the random nature of molecular motion, the resultant will change continuously in magnitude and in direction and at some instants may become large enough to effect a noticeable displacement of the particle in space. Since such fluctuations of the resultant (substantial deviations of it from zero) are random in nature the particle will move in space at random. When one observes the motion of such particles in a microscope, one sees them for the most part of the time to shift around one place, the speed of motion being the slower the larger the particle (Fig. 3.1).

The motion of minute particles bombarded by molecules moving at random is termed *Brownian motion*. It sort of copies molecular motion, but in a very slow way. Observations show Brownian motion to increase with the rise in temperature and to decrease with the drop in temperature.

The first to observe such motion was the Scottish botanist Robert Brown (1773-1858). In 1827 he discovered that pollen suspended in water shows a continuous random motion when viewed under a microscope. Brown was

Fig. 3.1 Brownian motion (points depict positions of particles at successive equal time intervals).



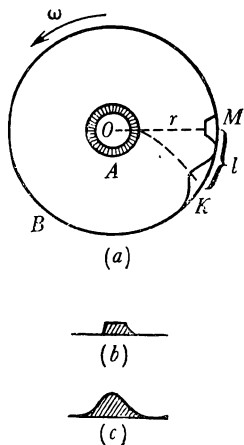
unable to establish the cause of such a motion of the particles.

Brownian motion is proof of the existence of random molecular motion in liquids and in gases; it is one of the most important phenomena in support of the kinetic theory.

### 3-3 Measuring Molecular Speeds

Studies of diffusion and of Brownian motion enable us to obtain some notions about the speed of random motion of molecules. One of the simpler and straightforward methods of assessing molecular speeds is the experiment carried out by Otto Stern (1888-1969) in 1920. The essence of the experiment is as follows.

**Fig. 3.2** Stern experiment: (a) setting viewed from above; (b) silver deposit at point  $M$  (cylinders at rest); (c) silver deposit at point  $K$  (cylinders rotating at angular speed  $\omega$ ).



Two hollow cylinders  $A$  and  $B$  are mounted at right angles to a horizontal platform that can rotate about an axis,  $O$  (Fig. 3.2a). The surface of  $B$  has no recesses while the surface of  $A$  has a narrow slit parallel to axis  $O$ . Axis  $O$  is a vertical silver-plated platinum wire connected to an electrical circuit. When a current flows in the wire, it is heated and silver evaporates from its surface. The silver atoms fly in all directions and condense mainly on the internal surface of cylinder  $A$ . Only a narrow beam of silver atoms passes through the slit and condenses in the region  $M$  of surface  $B$ . The width of the deposit on  $M$  is determined by the width of the slit in surface  $A$ . To prevent the scattering of silver atoms due to collisions with the molecules of air the entire setting is covered with a bell-jar and the air is pumped out of it. The less the width of the slit in surface  $A$  the narrower the deposit on  $M$  and the more accurate the determination of molecular speeds.

The measurement of the speed  $v$  itself is based on the following idea. If the whole setting is made to rotate at a constant angular speed  $\omega$ , the outer cylinder  $B$  will turn during the time the molecule flies from the slit to its surface and the deposit will shift from region  $M$  to a region  $K$ , the time the molecule flies along the radius  $r$  and the time point  $M$  on surface  $B$  shifts by a distance  $l = \widehat{KM}$  being equal. Since the molecule flies at a constant speed, it follows that

$$t = (r - r_A)/v \quad (3.1)$$

where  $v$  is the sought velocity, and  $r_A$  is the radius of surface  $A$  ( $r_A \ll r$ ). Since the linear velocity of the points on surface  $B$  is  $\omega r$ , time  $t$  can be expressed with the aid of another formula

$$t = l/\omega r \quad (3.2)$$

Hence

$$(r - r_A)/v = l/\omega r \quad (3.3)$$

Since  $\omega$ ,  $r$  and  $r_A$  remain constant in the course of the experiment and are determined beforehand, we can find  $v$  by measuring  $l$ . In Stern's experiment it proved to be close to 500 m/s.

Since the deposit on  $K$  is always blurred, one may infer that the silver atoms fly towards surface  $B$  at different speeds. Average velocities can be expressed mathematically with the aid of the formula

$$\bar{v} = \frac{v_1 + v_2 + \dots + v_n}{n} = \frac{1}{n} \sum_{i=1}^n v_i \quad (3.4)$$

We note that at 0 °C the average velocity of hydrogen molecules is 1840 m/s and that of nitrogen 493 m/s. The variation of the thickness of the deposit on  $K$  is an indication of the distribution of the molecules with respect to speed. A result that is inferred from it is that some molecules have velocities several times greater than the average.

(Ponder on the question where in Fig. 3.2a should be the trace of molecules with speeds in excess of the average speed  $\bar{v}$  and how will the position of the deposit change, if the current in the wire  $O$  is increased.)

### 3-4 Distribution of Molecular Speeds

A mass of gas occupying a constant volume with the pressure and the temperature remaining constant is said to be in a state of *thermal equilibrium*. The gas can remain in such a state for a long time provided the external conditions remain unaltered.

The prominent Scottish physicist James Clerk Maxwell (1831-1879) made theoretical studies of the random motion of the molecules of a gas in a state of equilibrium. The result was that molecular speeds should be quite different. In 1850 Maxwell found with the aid of the theory of probability a mathematical expression for the distribution of molecules of a gas in a state of equilibrium over the speeds of their random motion.

Let the total number of molecules be  $n$ . Denote the number of molecules whose speeds lie between  $v_1$  and  $v_2$  by  $\Delta n$ . In that case  $\Delta n/n$  is the fraction of the molecules whose velocities lie within the specified interval  $\Delta v = v_2 - v_1$ .

Clearly, if we take equal speed intervals  $\Delta v_1 = (405 - 400) \text{ m/s} = 5 \text{ m/s}$  and  $\Delta v_2 = (505 - 500) \text{ m/s} = 5 \text{ m/s}$ ,

say, for nitrogen molecules, the number of molecules inside those intervals will be different, because at a definite temperature some speeds are more common than the others.

The relative number of molecules whose speeds lie inside the interval  $\Delta v$  is proportional to the length of this interval and depends on the region where this interval was chosen:

$$\Delta n/n = y \Delta v \quad (3.5)$$

Here  $y$  depends on the magnitude of the speed of random motion, that is, it is a function of  $v$ , or  $y = f(v)$ .

Hence, formula (3.5) may be written in the form

$$\Delta n/n = f(v) \Delta v \quad (3.6)$$

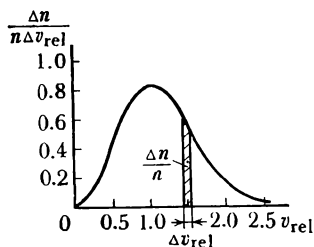
The function

$$f(v) = \Delta n/n\Delta v \quad (3.6a)$$

is termed the *distribution function of speeds for molecules in random motion*, or Maxwell's function. The mathematical expression for *Maxwell's function* is rather complicated; therefore, we present here only the graph of this function (Fig. 3.3). The speed corresponding to the maximum of Maxwell's function is termed the *most probable speed* and denoted  $v_p$ . The horizontal scale is a plot of the relative velocity of molecular motion,  $v_{rel} = v/v_p$ , so that the graph can be applied to different temperatures and gases.

The relative number of molecules inside a small specified interval of velocities  $\Delta v_{rel}$ , is expressed by the shaded area in Fig. 3.3. (Making use of Fig. 3.3, explain why the speed corresponding to the maximum of Maxwell's function is termed most probable and why the average velocity exceeds the most probable, and find the area bounded by the graph of Maxwell's function and the horizontal axis.)

Fig. 3.3 Maxwell's distribution curve for random motion of gas molecules.



### 3-5 Mass and Size of Molecules and Atoms

Many phenomena, for instance, Brownian motion, prove the dimensions of the molecules to be very small. There are numerous methods in physics and in chemistry that make possible an accurate determination of the dimensions of molecules, of atoms of chemical elements, and of particles of which they are made up. Some of those methods shall be discussed below.

The achievements of modern science make it possible to definitely establish both the dimensions and the masses of individual molecules. If we imagine the molecules as little balls, the diameters will in most cases prove to be less than

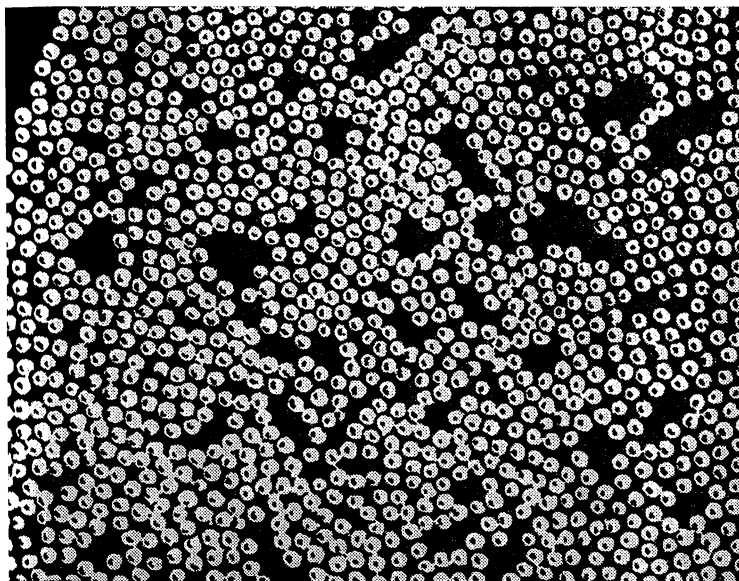


Fig. 3.4 Electron micrograph of DNA molecule.

a nanometer. For instance, the diameter of a water molecule is 0.26 nm (2.6 Å).

Presently various polymers are being widely used. Their molecules sometimes are made up of many thousands of atoms and their dimensions may be appreciably greater. This is true also of other organic substances. Photographs of some large molecules made with the aid of an electron microscope have been obtained (Fig. 3.4).

Molecules are so small that one can get an idea of their dimensions only by comparing them with other objects. A molecule of water is as small in comparison with a large apple as the apple is in comparison with the globe. The mass of a molecule of oxygen,  $O_2$  is  $53.5 \times 10^{-24}$  g and the mass of a hydrogen molecule,  $H_2$ , is  $3.34 \times 10^{-24}$  g. The mass of the lightest atom existing in nature, the hydrogen atom, H is  $1.672 \times 10^{-24}$  g.

To measure the masses of molecules and atoms in kilograms or grams proved to be inconvenient. For this reason an additional unit of measurement was introduced, the *atomic mass unit* (amu). The atomic mass unit is defined as one-twelfth of the atomic mass of the carbon isotope\*  $C^{12}$ .

\* An isotope is one of two or more atoms having the same atomic number but different atomic mass.

The mass of a molecule (an atom) expressed in atomic mass units is termed *relative molecular (atomic) mass*. It shows the number of times the molecular mass is greater than  $1/12$  of the mass of an atom of the  $C^{12}$  isotope. For instance, the relative mass of a hydrogen molecule is 2.015 94 and that of a nitrogen molecule 28.0134. Precise measurements have produced for the atomic mass unit the result  $1.660 \times 10^{-27}$  kg. Hence, knowing the molecular mass  $m_{\text{rel}}$  in atomic mass units one can find its mass in kilograms with the aid of the following formula

$$m = m_{\text{rel}} \times 1.660 \times 10^{-27} \text{ kg} \quad (3.7)$$

Note that relative atomic (molecular) mass is a dimensionless number.

### 3-6 Avogadro and Loschmidt Numbers

Suppose we have two gases: hydrogen and nitrogen. For the sake of simplicity we shall assume their relative molecular masses to be equal to 2 and 28, respectively. If we take 2 grams of hydrogen and 28 grams of nitrogen, the number of molecules in each mass will be the same. Indeed, the ratio of the masses of these gases is  $14 \div 1$  (same as the ratio of the masses of a molecule of nitrogen and a molecule of hydrogen). This means that in our example the number of the hydrogen and nitrogen molecules is the same.

It proved to be convenient for many practical applications to operate with amounts of substances containing equal numbers of molecules. For this reason in various calculations the concept of a *mole* is frequently made use of.

Mole (abbreviated *mol*) is the mass of a substance in grams numerically equal to its relative molecular mass. The example cited above shows that the number of molecules in a mole of any substance is the same. The number of molecules in a mole of a substance is termed the *Avogadro number* and denoted  $N_A$ .

If we imagine a substance with the molar mass equal precisely to unity, we would expect one mole of such a substance to have a mass of one gram. Since the mass of one molecule of such a substance is  $1.660 \times 10^{-27}$  kg, the Avogadro number can be found by the following procedure:

$$\begin{aligned} N_A &= \frac{10^{-3} \text{ kg per mole}}{1.660 \times 10^{-27} \text{ kg per molecule}} \\ &\approx 6.02 \times 10^{23} \text{ molecules per mole} \end{aligned}$$

In various calculations a substance may be considered as consisting not only of molecules but also of atoms, ions, electrons or other particles, that is, of specific structural elements. Therefore the term mole applies generally to the amount of substance (in the form of structural elements of a definite kind) the weight of which in grams is equal to the relative mass of the respective structural element. Obviously, the number of structural elements per mole will always be the same and equal to the Avogadro number.

We note that in the SI system the mole is one of the base units. Often the kilomole of substance is used: 1 kilomole = 1000 moles.

It was established in experiments aimed at investigating the properties of gases that a mole of any gas under standard conditions (0°C temperature and 1 atm pressure) occupies the volume  $22.4 \times 10^{-3} \text{ m}^3$ , or 22.4 litres. This result is in full agreement with Avogadro's law known from chemistry: equal volumes of all gases, under the same conditions of temperature and pressure, contain equal numbers of molecules.

The number of gas molecules per unit volume under standard conditions is termed the *Loschmidt number* and denoted  $N_L$ . It may easily be calculated by dividing the Avogadro number by the volume of a mole under standard conditions:

$$\begin{aligned} N_L &= \frac{6.02 \times 10^{23} \text{ molecules per mole}}{22.4 \times 10^{-3} \text{ m}^3 \text{ per mole}} \\ &\approx 2.7 \times 10^{25} \text{ molecules per m}^3 \end{aligned}$$

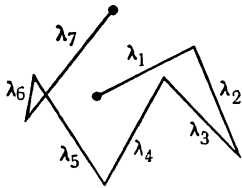
### 3-7 Mean Free Path

It was mentioned above that random motion of molecules in a gas results in numerous collisions. It was established that under standard conditions every gas molecule experiences on the average some  $10^9$  collisions with other molecules. We must note, however, that collisions must not be understood too literally, that is, not in the sense of collisions of solid bodies.

The proximity of colliding molecules depends on the mutual directions of their velocities and on the kinetic energy of their translational motion, that is, on temperature. Therefore the molecular diameters obtained in experiments depend on the temperature and, as such, serve only as an approximate characteristic of their dimensions. Because of that the numerical value of the molecular diameters



Fig. 3.5 Free paths of gas molecule.



determined in this fashion are termed *effective molecular diameters*.

The distance covered by a molecule between two successive collisions is termed free path and denoted  $\lambda$  (the Greek lambda). The free paths of a molecule between individual collisions may be quite different (Fig. 3.5). Therefore the mean free path  $\bar{\lambda}$  is introduced:

$$\bar{\lambda} = \frac{\lambda_1 + \lambda_2 + \dots + \lambda_z}{z} \quad (3.8)$$

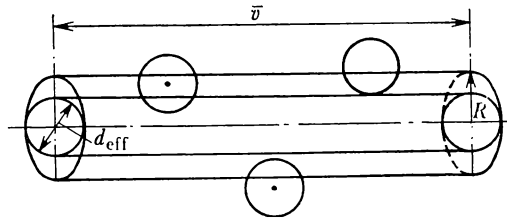
If  $\bar{z}$  denotes the average number of collisions of a gas molecule per second, then the sum in the numerator of (3.8) expresses the path covered by the molecule in one second, that is, the average velocity of its motion,  $\bar{v}$ . Hence,

$$\bar{\lambda} = \bar{v} / \bar{z} \quad (3.9)$$

Under standard conditions,  $\bar{\lambda}$  for air molecules is about  $10^{-7}$  m, or  $0.1 \mu\text{m}$ . Calculations show that under such conditions the molecules occupy 0.04 per cent of the volume. The remaining 99.96 per cent of the volume is free from molecules.

Let us make a rough estimate of the number of molecular collisions per second. Depict the path covered by the molecule in one second by a straight line (Fig. 3.6) of the length  $\bar{v}$ .

Fig. 3.6 Molecules collide if distance between their centres is less than effective molecular diameter.



Suppose the number of molecules per unit volume in the surrounding space is  $n_0$ . In that case our molecule moving along a rectilinear path will strike all molecules whose centres lie inside a cylinder of radius  $R$  equal to the effective molecular diameter  $d_{\text{eff}}$ .

Since the volume of this cylinder is  $\pi R^2 \bar{v} = \pi d_{\text{eff}}^2 \bar{v}$ ; altogether there will be  $\pi d_{\text{eff}}^2 \bar{v} n_0$  molecules in it. Those molecules will be involved in collisions in the course of one second. Hence,  $\bar{z} = \pi d_{\text{eff}}^2 \bar{v} n_0$ . A more accurate estimate yields

$$\bar{z} = \sqrt{2} \pi d_{\text{eff}}^2 \bar{v} n_0 \quad (3.10)$$

We assumed our molecule to move with velocity  $\bar{v}$  and the other molecules in its way to be at rest. Actually they, too, move with the average velocity  $\bar{v}$ . Therefore we should take for our molecule its velocity relative to other molecules,  $\bar{v}_{rel}$ .

The angles between the velocity vectors of the colliding molecules may be quite different from 0 to 180 degrees; for the average angle of 90 degrees,  $\bar{v}_{rel} = \sqrt{\bar{v}^2 + \bar{v}^2} = \bar{v}\sqrt{2}$ .

Substituting the value for  $\bar{z}$  in (3.10) into (3.9), we have

$$\bar{\lambda} = \frac{1}{\sqrt{2} \pi d_{eff}^2 n_0} \quad (3.11)$$

### 3-8 Gaseous Pressure. Pressure Gauges

Gas molecules striking against the surface of a body exercise pressure on it. This pressure is the greater the greater the average kinetic energy of translational motion of the gas molecules and their number per unit volume. Recall that pressure is measured by the normal force acting on a unit surface:

$$p = F_n/A \quad (3.12)$$

We derive the SI unit of pressure:

$$p = \frac{1 \text{ N}}{1 \text{ m}^2} = 1 \frac{\text{N}}{\text{m}^2} = 1 \frac{\text{kg} \cdot \text{m}/\text{s}^2}{\text{m}^2} = 1 \frac{\text{kg}}{\text{m} \cdot \text{s}^2} = 1 \text{ pascal (Pa)}$$

Thus, the unit of pressure in the SI system is the *pascal*, which is the pressure that a force of 1 N exerts on an area of 1 m<sup>2</sup>.

In practice other units of pressure are also used. Some of them are listed below.

The technical atmosphere (abbreviated at) is the pressure that a force of 1 kgf (kilogram-force) exerts on an area of 1 cm<sup>2</sup>:

$$1 \text{ at} = 1 \text{ kgf/cm}^2 \approx 9.81 \times 10^4 \text{ N/m}^2 = 9.81 \times 10^4 \text{ Pa}$$

The millimetre of mercury (abbreviated mmHg) is the pressure exerted by a column of mercury 1 mm high on a horizontal surface. Since the formula for the pressure at the depth  $h$  in a liquid due to its weight is

$$p = \rho gh \quad (3.13)$$

we can easily establish the following relation:

$$1 \text{ mmHg} = 13.6 \times 10^3 \text{ kg/m}^3 \times 9.81 \text{ m/s}^2 \times 10^{-3} \text{ m} \approx 133 \text{ Pa}$$

Fig. 3.7 Metal pressure gauge.

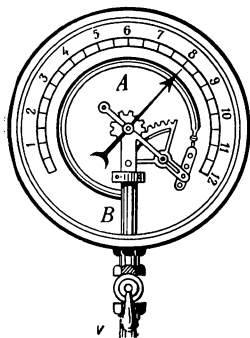


Fig. 3.8 Closed-tube liquid manometer.

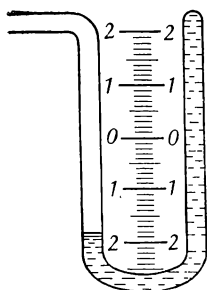
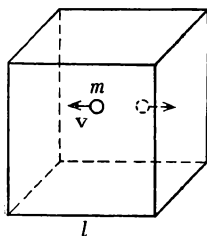


Fig. 3.9 The molecule striking left wall strikes it again after travelling distance  $2l$  at velocity  $v$ .



The physical atmosphere (abbreviated atm) is the pressure that a column of mercury 760 mm high exerts on a horizontal surface:

$$1 \text{ atm} = 1.033 \text{ at} = 1.013 \times 10^5 \text{ Pa}$$

This quantity is accepted as the standard atmospheric pressure.

One bar (bar) is the pressure equal to  $10^5$  Pa:

$$1 \text{ bar} = 10^5 \text{ Pa}$$

Instruments for measuring pressures are termed pressure gauges and manometers. High pressures are measured with the aid of the *metal pressure gauge* (Fig. 3.7). Its principal part is a bent metal tube *A*. Its open end is soldered to the tube *B* and the closed end is connected to a pointer. When valve *V* is opened, the gas enters tube *A* and straightens it out. This is because the area of the concave internal surface of the tube exceeds that of the convex surface. But the pressure of the gas is the same in all directions. This means that the force of pressure exerted on the concave surface exceeds that on the convex surface, the net force straightens out tube *A* so that its end moves the instrument's pointer across the scale.

The instrument used to measure very low pressures is the *closed-tube liquid manometer* (Fig. 3.8). If its open end is connected to the vessel in which the gas pressure is low, the level of the liquid in the closed tube will sink appreciably. The pressure in the vessel is found from the difference of levels in the open and closed tubes.

### 3-9 Kinetic Calculation of the Pressure

The pressure of a gas in a vessel is because the molecules strike the walls of the vessel. Let us by using the kinetic theory, find the relation between the pressure of a gas and other quantities that characterize it.

Consider a hollow cubical vessel  $l$  cm on a side. Suppose each unit of volume inside the cube contains  $n_0$  identical molecules of a gas. Since the molecules are in motion, each possesses a momentum  $m\mathbf{v}$ . (Here  $m$  is the mass of a molecule and  $\mathbf{v}$  its velocity.) Assume that the impacts of the molecules on the walls are elastic. In that case, if a molecule moves at right angles to the wall of the vessel (Fig. 3.9), in striking the wall it stops, that is, loses its momentum  $m\mathbf{v}$ , and then recoils from the wall and moves in the opposite

direction with a velocity  $-\mathbf{v}$ , that is, receives a momentum  $m\mathbf{v}$  of opposite direction. Hence, in the course of its impact the momentum of such a molecule changes in magnitude by  $2mv$  and the wall receives a momentum of the same magnitude,  $2mv$ .

Let this molecule move unimpeded between the left-hand and right-hand walls of the vessel. With each impact it imparts to the wall a momentum  $2mv$ . In accordance with the laws of mechanics the average force with which the molecule acts on the wall is

$$F_1 = 2mv/\Delta t$$

where  $\Delta t$  is the time interval in which the molecule moves from one wall to another and back, that is,  $\Delta t = 2l/v$ , since this is the time interval between successive impacts of the molecule against the wall. Therefore

$$F_1 = 2mv \times v/2l = mv^2/l$$

The force of pressure of the gas on the vessel's wall is equal to the sum of the forces of impact of the individual molecules against this wall:

$$\begin{aligned} F_n &= F_1 + F_2 + \dots + F_N = \frac{mv_1^2}{l} + \frac{mv_2^2}{l} + \dots + \frac{mv_N^2}{l} \\ &= \frac{m}{l} (v_1^2 + v_2^2 + \dots + v_N^2) = \frac{m}{l} v_{\text{rms}}^2 N \end{aligned}$$

where  $N$  is the total number of molecules flying between two opposite walls of the vessel, and  $v_{\text{rms}}$  is the root-mean-square speed of the molecules (see (2.2) in Section 2-5). Since the number of molecules in a unit volume is  $n_0$ , the total number of molecules inside the cube is  $n_0 l^3$ .

The molecules move absolutely at random so that all directions of motion are equally probable for any one of them. Therefore it may be assumed that the number of molecules moving between two opposite walls at right angles to them is one-third of the total, or  $N = n_0 l^3/3$ . Hence

$$\begin{aligned} F_n &= \frac{m}{l} v_{\text{rms}}^2 N = \frac{mv_{\text{rms}}^2}{l} \frac{1}{3} n_0 l^3 \\ &= \frac{1}{3} n_0 m v_{\text{rms}}^2 l^2 \end{aligned}$$

Since  $p = F_n/A$ , we obtain

$$p = \frac{1}{3} n_0 m v_{\text{rms}}^2 = \frac{2}{3} n_0 \frac{mv_{\text{rms}}^2}{2}$$

Recalling that  $mv_{\text{rms}}^2/2 = \bar{K}_{\text{trans}}$ , we obtain finally

$$p = \frac{2}{3} n_0 \bar{K}_{\text{trans}} \quad (3.14)$$

Now it is obvious that the pressure of a gas is proportional to the average kinetic energy of translational motion of its molecules and to their number per unit volume.

Formula (3.14) is very important and is termed the *principal equation of the kinetic theory of gases*.

### 3-10 Vacuum

The number of molecules per unit volume of air close to the Earth's surface is represented by the Loschmidt number  $N_L$  (see Section 3-6).

If in a space containing gas the number of gas molecules per unit volume  $n_0$  is below the Loschmidt number  $N_L$ , this space is said to contain rarefied gas. If the pressure in the space is considerably below the atmospheric (a thousand times or more), the space is said to be a vacuum. The higher the rarefaction of the gas in a closed space, that is, the lower the concentration  $n_0$  of its molecules, the higher the vacuum.

It follows from formulae (3.10) and (3.11) that a decrease in  $n_0$  leads to a decrease in the number of molecular collisions  $\bar{z}$  and to an increase in the mean free path  $\bar{\lambda}$ . Therefore one can imagine such a high rarefaction (such a small concentration  $n_0$ ) that the mean free path  $\bar{\lambda}$  will become equal to the dimensions of the vessel containing the gas. Obviously, any further rarefaction will not change  $\bar{\lambda}$ . The term *high vacuum* means that the mean free path of the molecules in the vessel is determined solely by its dimensions. In this case the molecules for the most part fly unimpeded between the walls of the vessel and only after many impacts against the walls they occasionally collide with each other.

It follows from formula (3.14) that the higher the vacuum in a vessel containing gas the smaller the gas pressure  $p$ . If high vacuum is attained in, say, a 3-l vessel, the number of gas molecules per cubic centimetre will still be about  $10^{12}$ , the pressure being approximately  $10^{-4}$  mmHg.

Modern technology is capable of producing a vacuum with a gas pressure below  $10^{-11}$  mmHg, but there is still some hundreds of thousands of gas molecules per cubic centimetre.

High vacuum is required for the operation of many scientific and technological devices, for instance, TV tubes, X-ray tubes, radio tubes, etc.

The maximum vacuum in nature may be obtained in the absence of molecules or any other particles (the Latin word *vacuus* means empty). However, space devoid of particles cannot be considered to be empty since there are always gravitational and electromagnetic fields in it.

Matter in interstellar space is in an extremely rarefied state. It consists of the minutest particles of "dust", atoms and gas molecules, mainly hydrogen. Gas and "dust" are dispersed throughout the interstellar space, but their distribution is not uniform. They form extensive clouds of irregular shape called gaseous and "dust" nebulae. The concentration of particles in these nebulae may be as high as several dozen particles per cubic centimetre, which is still several tens of thousands of times lower than in the highest vacuum obtained on Earth.

## The Ideal Gas

## 4

### 4-1 Properties of Ideal Gas

It is practically impossible in studying natural and technological phenomena to take into account all the factors influencing a specific phenomenon. However, the most important factors can always be established from experiments. Then one shall be able to neglect all the other factors of a lesser importance and thus obtain an ideal (simplified) notion about the phenomenon. The next step is the theoretical evaluation of the course of the phenomenon under idealized conditions, that is, when only the most important factors are in operation. The model construed in this fashion is useful, because it helps in the study of real processes and makes it possible to predict their course under various conditions. Let us consider one of such idealized concepts.

As was mentioned above (see Section 3-1), the physical properties of a gas are determined by the random motion of its molecules, the molecular interaction little affecting its properties and the interaction itself having the nature of collisions in which the molecular attraction can usually be neglected. The forces of repulsion act only during very short time intervals and the gas molecules most of the time move as free particles. This makes it possible to introduce the concept of *ideal gas*; in such a gas there are no forces of attraction between the molecules at all and the interaction of molecules can often be totally neglected so that the mole-

cules may be considered to be quite free. If such a model of a gas is applicable to real gases, their properties should not depend noticeably on their nature. At pressures not too high and temperatures not too low this is indeed the case.

As we know, the volume of the gas molecules themselves constitutes a tiny fraction of the volume occupied by the gas (if it is not compressed). This, by the way, also supports the contention that the nature of gas molecules has little effect on its properties. Therefore, the volume of the molecules of ideal gas must always be negligible as compared with the volume occupied by the gas. In this sense the molecules may be imagined as particles and the molecules of a multiatomic gas as rigidly connected particles.

To sum up, ideal gas is understood to be a gas in which the interaction of the molecules can be neglected and the molecules themselves can be regarded as particles. Since the molecules of ideal gas are never attracted to one another, it should remain in the gaseous state under all external conditions.

The concept of an ideal gas is a useful one, because all real gases at small pressures and at temperatures not too low obey simple general laws true, strictly speaking, only for an ideal gas. Theoretical studies of properties of the ideal gas enabled some conclusions extending our knowledge of natural phenomena to be drawn.

At high pressures the molecules of a real gas are so close that the forces of attraction begin to play a noticeable part. The volume of the molecules, too, exerts considerable influence on their behaviour in those conditions. Therefore at high pressures the properties of real gases depend on their nature, and they considerably differ from the properties of the ideal gas. The same applies to real gases at low temperatures. On account of their properties, hydrogen and helium are closer to the ideal gas than other gases.

#### **4-2 Change of Gaseous Pressure with Temperature at Constant Volume**

Let us discuss the dependence of the pressure of a gas on temperature when its mass and volume remain constant.

Take a closed vessel containing the gas and start heating it (Fig. 4.1). To measure the temperature  $t$  use a thermometer and to measure the pressure a pressure gauge, G. First place the vessel into melting snow and denote the gas pressure at  $0^{\circ}\text{C}$  by  $p_0$ ; next start heating the external vessel gradually, recording the values of  $p$  and  $t$  for the

gas. The  $p$  versus  $t$  plot obtained in this experiment will be a straight line (Fig. 4.2a). Should this plot be extrapolated to the left, it would intersect the  $t$  axis in point  $A$  corresponding to zero pressure of the gas.

The triangles in Fig. 4.2a are similar, which yields

$$\frac{p_0}{OA} = \frac{\Delta p}{\Delta t}, \quad \text{or} \quad \frac{1}{OA} = \frac{\Delta p}{p_0 \Delta t} \quad (4.1)$$

Denoting the constant  $1/OA$  by  $\gamma$ , we obtain

$$\gamma = \frac{\Delta p}{p_0 \Delta t} \quad (4.2)$$

$$p = \gamma p_0 \Delta t \quad (4.2a)$$

The proportionality factor  $\gamma$  must express the dependence of the variation of gas pressure on its nature.

The quantity  $\gamma$  which characterizes the dependence of pressure variation in the course of temperature variation at constant volume and constant mass of the gas on its nature is termed the *pressure coefficient* of the gas. The pressure coefficient is the fractional change in pressure when the temperature of the gas increases from  $0^\circ\text{C}$  to  $1^\circ\text{C}$  (see Eq. (4.2)) at constant volume.

Let us derive a unit for measuring  $\gamma$ :

$$\gamma = \frac{1 \text{ Pa}}{1 \text{ Pa} \times 1^\circ\text{C}} = 1^\circ\text{C}^{-1}$$

Repeating the experiment described above for various gases and for different masses, one can establish the fact that within experimental errors the position of point  $A$  of various plots will always be the same (Fig. 4.2b).

The length of the section  $OA$  will be equal to  $273^\circ\text{C}$ . Consequently, in all cases the temperature at which the pressure of a gas should turn zero is the same and equal to  $-273^\circ\text{C}$ , the value of the pressure coefficient being  $\gamma = 1/OA = 1/273^\circ\text{C}^{-1}$ . Note that the precise value is  $1/273.15^\circ\text{C}^{-1}$ . In solving problems the approximate value  $1/273^\circ\text{C}^{-1}$  is mainly used.

The French physicist Jaques A. C. Charles (1746-1823) was the first to obtain experimental values of  $\gamma$ . In 1787 he established (but never published) the following law: the pressure coefficient is independent of the nature of the gas and is equal to  $1/273.15^\circ\text{C}^{-1}$ . We know now that this is true only for gases of small density and for small temperature variations. It should also be noted that only the ideal gas strictly obeys this law.

Fig. 4.1 Heating gas at constant volume.

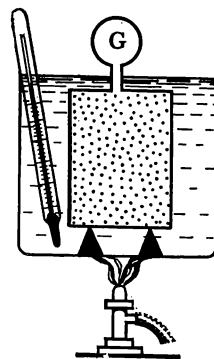
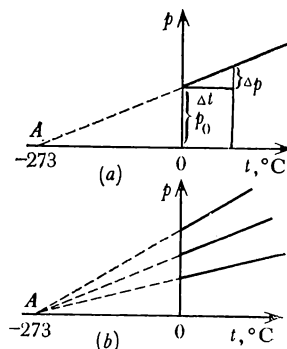


Fig. 4.2 (a) Variation of pressure as function of temperature at constant volume; (b) for different gases points of intersection of plots with  $t$  axis coincide.





Let us find how one can determine the pressure of a gas,  $p_t$  at any temperature  $t$ . It may be seen from Fig. 4.2a that

$$\Delta p = p_t - p_0 \quad \text{and} \quad \Delta t = t - 0 = t$$

Substituting  $\Delta p$  and  $\Delta t$  into Eq. (4.2), we obtain

$$p_t - p_0 = \gamma p_0 t, \quad \text{or} \quad p_t = p_0 (1 + \gamma t) \quad (4.3)$$

Since  $\gamma \approx 1/(273^\circ\text{C})$ , Eq. (4.3) may be used for solving problems in the form

$$p_t = p_0 \left( 1 + \frac{1}{273^\circ\text{C}} t \right) \quad (4.4)$$

### 4-3 Absolute Zero

It was noted in the preceding section that point  $A$  in Fig. 4.2b corresponds to  $p = 0$ . Let us find now under what conditions the pressure of an ideal gas will be zero.

Since gaseous pressure is due to the impacts of the molecules moving at random, the decrease in pressure of a gas being cooled is explained by the decrease in the average energy of translational motion of the gas molecules,  $\bar{K}_{\text{trans}}$ . The pressure of the gas will turn zero when the energy of translational motion of its molecules turns zero. This follows also from formula (3.14):

$$p = \frac{2}{3} n_0 \bar{K}_{\text{trans}}$$

(in experiments described in Section 4-2,  $n_0$  remains constant and does not turn zero). It is now obvious that the process of cooling a gas should have a limit corresponding to the absence of translational motion of its molecules.

The temperature at which the translational motion of the molecules should cease is termed *absolute zero*. Note that in nature there cannot be a temperature below absolute zero. Indeed, at a temperature below absolute zero the energy of translational motion of molecules must be negative and this is impossible.

Let us find now at what temperature the molecules of an ideal gas should stop. Since the ideal gas remains in the gaseous state at all temperatures, Eq. (4.4) is valid for absolute zero. Therefore

$$0 = p_0 \left( 1 + \frac{1}{273^\circ\text{C}} t \right)$$

Since  $p_0 \neq 0$ , we have

$$1 + \frac{1}{273^\circ\text{C}} t = 0, \quad \text{or} \quad t = -273^\circ\text{C}$$

Remember that the same value of  $t$  can be obtained from the plots in Fig. 4.2*b*. The more precise value of the absolute zero is  $-273.15^{\circ}\text{C}$ .

The eminent British scientist Sir William Thomson (Lord Kelvin; 1824-1907) suggested that the value of absolute zero obtained above corresponds to the cessation of translational motion of molecules of all substances. It follows from theoretical considerations that no body can be cooled to absolute zero. However, it is possible to attain temperatures very close to absolute zero. In physical laboratories temperatures have been obtained only some thousandths of a degree above absolute zero.

Note that only thermal motion of molecules (or atoms) stops when the absolute zero is approached, but by no means all types of motion—elementary particles inside atoms continue to move.

#### 4-4 Thermodynamic Temperature Scale

We already know that  $0^{\circ}\text{C}$  is conventionally accepted as the temperature of melting ice at standard pressure and  $100^{\circ}\text{C}$  as the temperature of boiling water at standard pressure. A hundredth of this temperature interval is the practical unit for measuring temperature, one degree Celsius ( $^{\circ}\text{C}$ ). However, the readings of alcohol and mercury thermometers whose scales from  $0^{\circ}\text{C}$  to  $100^{\circ}\text{C}$  consist of 100 equal divisions coincide only at  $0^{\circ}\text{C}$  and  $100^{\circ}\text{C}$ . This indicates that the thermal expansion of at least one of those substances is not uniform and one cannot obtain a unique temperature scale with the aid of such thermometers.

To construct a universal temperature scale one should have a quantity the variation of which with temperature would be independent of the kind of substance used in the thermometer. The pressure of a gas may serve as such a quantity, for the pressure coefficient of not very dense gases is independent of the nature of the gas and equal to that of the ideal gas. The best substance for thermometry would be ideal gas. Since the properties of rarefied hydrogen are the closest to those of the ideal gas, the hydrogen (ideal-gas) thermometer is the best for measuring temperatures. It consists of a tank containing rarefied hydrogen and connected to a sensitive pressure gauge. Since the pressure and temperature of hydrogen are interrelated by Eq. (4.3), one may find the temperature from the readings of the pressure gauge.

The International Practical Temperature Scale (IPTS 68) consists of a temperature scale established with the aid of

a hydrogen thermometer whose  $0^\circ$  corresponds to the temperature of the melting ice and  $100^\circ$  to that of boiling water. This scale is sometimes called Celsius scale.

Note that the zero on the Celsius scale is a matter of convention, and so is the value of a degree Celsius. This means that from a scientific point of view other definitions of a temperature scale is also feasible.

A sensible choice of the temperature scale makes it possible to simplify formulae and to gain better understanding of the physical meaning of the regularities observed. With this in view, W. Thomson suggested that a new temperature scale be introduced, the modern term for which is the *thermodynamic temperature scale*. Sometimes it is termed the *Kelvin scale*. On the scale the origin corresponds to absolute zero and the value of a degree is made to coincide as exactly as possible with one degree Celsius.

In the SI system the unit for measuring temperatures termed *kelvin* (K) is one of the basic units, and the thermodynamic temperature scale is used to record temperatures.

In accordance with an international convention the value of the kelvin is defined as follows: the temperature of the triple point of water (see Section 14.8) is accepted as being equal precisely to 273.16 K. Therefore to obtain the value of the kelvin the temperature interval between absolute zero and the triple point of water measured with the aid of a hydrogen (ideal-gas) thermometer should be divided into 273.16 parts.

Since the temperature corresponding to the triple point of water on the Celsius scale is  $0.01^\circ\text{C}$ , the temperature of melting ice on the new scale will be 273.15 K. And since the value of a kelvin is equal to that of a degree Celsius, the temperature of boiling water at standard pressure will be 373.15 K. For the sake of simplicity we shall assume below the temperatures of melting ice and of boiling water to be 273 K and 373 K, respectively.

The temperature expressed in kelvins on the thermodynamic scale is termed *thermodynamic*, or *absolute*, and is denoted  $T$ . The relation between absolute temperature and centigrade temperature is

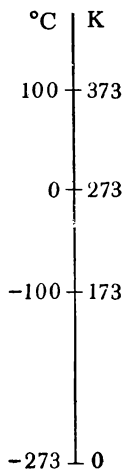
$$T = 273.15^\circ\text{C} + t \quad (4.5)$$

Usually the approximate formula is employed:

$$T = 273^\circ\text{C} + t \quad (4.6)$$

Figure 4.3 schematically shows this relation. It follows from the diagram that under no conditions can absolute temperature be negative.

Fig. 4.3 Schematic representation of equal temperatures in degrees Celsius and kelvins.



#### 4-5 Relation of Temperature to Kinetic Energy of Gas Molecules

The  $p$  versus  $t$  dependence established by J. Charles is depicted in Fig. 4.2a. If the origin is placed in point A, the straight line will pass through it. All the values of temperature will rise by  $273^\circ\text{C}$  because the length of the section  $OA$  is  $273^\circ\text{C}$  (see Section 4-2). This agrees with formula (4.6),  $T = 273^\circ\text{C} + t$ , since now the  $t$  axis is marked in values of absolute temperature (Fig. 4.4). In this case

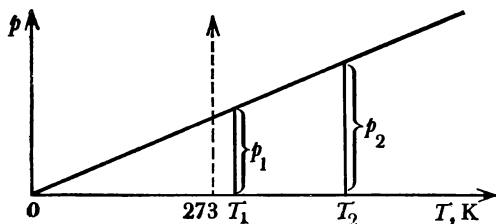


Fig. 4.4 Variation of ideal gas pressure as function of absolute temperature at constant volume.

there will be a direct proportionality between  $p$  and  $T$ . Indeed, from the similarity of the triangles in Fig. 4.4 we obtain

$$p_1/T_1 = p_2/T_2 \quad (4.7)$$

Hence, on the one hand, the pressure of a gas is directly proportional to its absolute temperature and, on the other, (see (3.15)), it is directly proportional to the average kinetic energy of the translational motion of gas molecules (with the mass and volume of the gas being constant). This means that  $\bar{K}_{\text{trans}}$  is directly proportional to the absolute temperature of the gas,  $T$ . The eminent Austrian physicist Ludwig Boltzmann (1844-1906) proposed to write the proportionality factor in the  $\bar{K}_{\text{trans}}$  versus  $T$  dependence in the form  $(3/2) k_B$ , where  $k_B$  is a constant termed the *Boltzmann constant*:

$$\bar{K}_{\text{trans}} = \frac{3}{2} k_B T \quad (4.8)$$

The numerical value of the Boltzmann constant in the SI system is (see Section 5-3)

$$k_B = 1.38 \times 10^{-23} \text{ J/K}$$

It follows from formula (4.8) that the average kinetic energy of translational motion of molecules is independent

of the nature of a gas and is determined entirely by its temperature. If we substitute  $\bar{K}_{\text{trans}}$  from (4.8) into (3.14), we get

$$p = n_0 k_B T \quad (4.9)$$

We see that the pressure of a gas does not depend on its nature and is determined only by the concentration of molecules,  $n_0$ , and the temperature of the gas,  $T$ .

Since at a specified temperature the average kinetic energy of translational motion of molecules is the same for various gases, we can write

$$\frac{1}{2} m_1 v_{\text{rms } 1}^2 = \frac{1}{2} m_2 v_{\text{rms } 2}^2$$

which yields

$$\frac{v_{\text{rms } 1}}{v_{\text{rms } 2}} = \frac{\sqrt{m_2}}{\sqrt{m_1}} \quad (4.10)$$

At a fixed temperature the root-mean-square speeds of molecular motion are inversely proportional to the square roots of the molecular masses.

## Ideal - Gas Equation of State

### 5-1 Thermodynamic Properties

It was demonstrated in the preceding chapters that quantities characteristic of the world of molecules, for instance, the energy of a molecule, its velocity, its mass, etc., can be used to describe the properties of a gas. We can determine the numerical values of these quantities only from calculations. All such quantities are said to be *microscopic* (from the Greek *mikros* for small).

However, one may also describe the properties of gases using such quantities whose numerical values can be determined by simple measurement with the aid of instruments, as, for instance, pressure, temperature and volume of a gas. The values of such quantities are the result of combined action of great numbers of molecules and for this reason they are termed *macroscopic* (from the Greek *makros* for great).

Relation (3.14),  $p = (2/3)n_0\bar{K}_{\text{trans}}$ , establishes the relation between the microscopic and macroscopic parameters

of a gas. For this reason formula (3.14) is sometimes called the principal equation of the kinetic theory of gases. The macroscopic quantities that provide a unique description of the state of a gas are termed *thermodynamic properties* of a gas. The most important thermodynamic properties of a gas are its volume  $V$ , pressure  $p$  and temperature  $T$ .

A specified mass of a gas  $m$  at constant  $p$ ,  $V$  and  $T$  will be in a state of equilibrium. The changes in those parameters mean that a process takes place in the gas. If this process can be represented as a continuous succession of equilibrium states of the gas, it is termed an *equilibrium process*. An equilibrium process should proceed at a sufficiently slow rate, since the pressure and the temperature cannot be the same throughout the volume if the parameters change rapidly. This chapter deals only with equilibrium processes in gases in the course of which the mass of the gas remains constant.

At the end of a process the gas will be in a new state with its parameters assuming new numerical values generally different from the respective values at the start of the process. If the values of all the parameters of a constant mass of gas turn out to be the same at the start and at the end of the process, the process is said to be *cyclic*, or *closed*.

The relation between the values of specific parameters at the start and the end of a process is termed the *gas law*. The gas law establishing the correspondence between all the three gas parameters is termed the *combined gas law*.

Note in addition that there is no such process that brings about the change of only one of the gas parameters, since the values of all the parameters are interrelated. An example of this contention is the law that expresses the relation between two parameters, pressure and temperature.

## 5-2 Combined Gas Law

The relation between the pressure, volume and temperature of a given mass of a gas is expressed by formula (4.9):

$$p = n_0 k_B T$$

We note that  $n_0$  denotes the number of molecules per unit volume of gas,  $n_0 = N/V$ , where  $N$  is the total number of molecules and  $V$  the volume of the gas. Then

$$p = \frac{N}{V} k_B T, \quad \text{or} \quad \frac{pV}{T} = N k_B \quad (5.1)$$

Since for a constant mass of the gas  $N$  remains constant and since  $k_B$  is a universal constant, the right-hand side

of (5.1) is a constant:

$$pV/T = \text{constant} \quad (5.2)$$

Since the values of  $p$ ,  $V$  and  $T$  in (5.2) refer to the same state of the gas, we can formulate the combined gas law as follows: for a constant mass of gas the product of its volume, its pressure and its reciprocal absolute temperature remains constant for all states of this mass of gas.

Hence, if we introduce the notation of  $p_1$ ,  $V_1$  and  $T_1$  for the numerical values of the parameters at the start of a process involving a specified mass of gas, and the notation  $p_2$ ,  $V_2$  and  $T_2$  for the same parameters at the end of the process, we will obtain

$$p_1 V_1 / T_1 = p_2 V_2 / T_2 \quad (5.3)$$

Formulae (5.2) and (5.3) represent the mathematical expression of the combined gas law.

In practice one sometimes has to find the volume  $V_0$  of a given mass of gas under standard conditions, that is, at  $T_0 = 273$  K and  $p_0 = 1.013 \times 10^5$  Pa. Denoting the values of parameters of this mass of gas in a state other than the standard by  $p$ ,  $V$  and  $T$ , we obtain according to (5.3)

$$V_0 p_0 / T_0 = V p / T, \text{ or } = V_0 \frac{p T_0}{p_0 T} \quad (5.4)$$

Formula (5.4) makes possible the reduction of a given mass of a gas to standard conditions.

### 5-3 Universal Gas Constant

Formula (5.1) is valid for an arbitrary mass of a gas containing  $N$  molecules. If we apply this formula to a mole of a gas, we must substitute the Avogadro number for  $N$  and the volume of a mole of that gas  $V_{\text{mole}}$  for  $V$ :

$$p V_{\text{mole}} / T = N_A k_B$$

Since one mole of any gas contains a definite number of molecules,  $N_A$ , the product  $N_A k_B$  will be the same for all gases, that is, it will be independent of the nature of the gas. The designation for  $N_A k_B$  is  $R$  and the term for it is the *universal gas constant*. Hence

$$p V_{\text{mole}} / T = R \quad (5.5)$$

where

$$R = N_A k_B \quad (5.6)$$

The numerical value of  $R$  may be found by applying (5.5) to one mole of gas under standard conditions, since in this case  $V_{\text{mole}} = 22.4 \times 10^{-3} \text{ m}^3/\text{mole}$  (see Section 3-6). Indeed,

$$R = \frac{pV_{\text{mole}}}{T} = \frac{1.013 \times 10^5 \text{ N/m}^2 \times 22.4 \times 10^{-3} \text{ m}^3/\text{mol}}{273 \text{ K}} = 8.31 \frac{\text{N} \cdot \text{m}}{\text{mol} \cdot \text{K}}$$

that is

$$R = 8.31 \text{ J}/(\text{mol} \cdot \text{K})$$

This numerical value of  $R$  in the SI system should be remembered because it is frequently used in calculations and in solving problems.

Now we may easily find the numerical value of the Boltzmann constant. We obtain from (5.6)  $k_B = R/N_A$ . Substituting numerical values of  $R$  and  $N_A$ , we compute  $k_B$ :

$$\begin{aligned} k_B &= \frac{8.31 \text{ J}/(\text{mol} \cdot \text{K})}{6.02 \times 10^{23} \text{ molecules per mole}} \\ &= 1.38 \times 10^{-23} \frac{\text{J}}{\text{K}} \text{ per molecule} \end{aligned}$$

#### 5-4 The Ideal-Gas Law

Let us see what form expression (5.1) will assume if the universal gas constant,  $R$ , is introduced in it. Since  $N$  is the total number of molecules in a mass of gas and  $N_A$  is the number of molecules in a mole, we have

$$N = \nu N_A$$

where  $\nu$  is the number of moles in the mass of gas  $m$ . Therefore

$$pV/T = \nu N_A k_B$$

Since  $N_A k_B = R$  and  $\nu$  is equal to the mass of gas divided by the mass of a mole of that gas  $\mu$ , we obtain

$$\frac{pV}{T} = \frac{m}{\mu} R, \quad \text{or} \quad pV = \frac{m}{\mu} RT \quad (5.7)$$

Equation (5.7) is termed the *ideal-gas law*, or the ideal-gas equation of state for an arbitrary mass. For a mole of gas the ideal-gas law assumes the form

$$pV_{\text{mole}} = RT \quad (5.8)$$



Using formula (5.7), we can easily find the quantities that determine the density of a gas. Since  $\rho = m/V$ , Eq. (5.7) yields

$$\rho = p\mu/RT \quad (5.9)$$

### 5-5 Dependence of Root-Mean-Square Speed of Gas Molecules on Temperature

Let us now find how one can calculate the root-mean-square speed of molecular motion,  $v_{\text{rms}}$ . Since the average kinetic energy of translational motion of molecules  $\bar{K}_{\text{trans}}$  is equal to  $(3/2) k_B T$ , we can write

$$\frac{1}{2} m v_{\text{rms}}^2 = \frac{3}{2} k_B T, \quad \text{or} \quad v_{\text{rms}} = \sqrt{\frac{3k_B T}{m}} \quad (5.10)$$

Note that  $m$  in (5.10) is the mass of a molecule in kilograms. Since  $k_B = R/N_A$ , we obtain

$$v_{\text{rms}} = \sqrt{\frac{3RT}{mN_A}}$$

Since  $mN_A$  is the mass of one mole,  $\mu$  (see Section 3-6) we have

$$v_{\text{rms}} = \sqrt{3RT/\mu} \quad (5.11)$$

Finally, it follows from (5.9) that

$$RT/\mu = p/\rho$$

Therefore

$$v_{\text{rms}} = \sqrt{3p/\rho} \quad (5.12)$$

The root-mean-square speed can be found from any of the formulae (5.10), (5.11), or (5.12). Formulae for the arithmetic mean speed and for the most probable speed may be obtained from the Maxwell function. The arithmetic mean speed is

$$v = \sqrt{\frac{8k_B T}{\pi m}} = \sqrt{\frac{8RT}{\pi \mu}} = \sqrt{\frac{8p}{\pi \rho}} \quad (5.13)$$

and the most probable speed is

$$v_p = \sqrt{\frac{2k_B T}{m}} = \sqrt{\frac{2RT}{\mu}} = \sqrt{\frac{2p}{\rho}} \quad (5.14)$$

(making use of the plot of the Maxwell function (Fig. 3.3) explain why  $v_p < \bar{v} < v_{\text{rms}}$ ).

## 5-6 Isochoric Process

Processes in which the mass of the gas and one of the parameters of state remain constant are termed *isoprocesses* (from the Greek *isos* for equal, identical).

Since there are three state properties, there are three different isoprocesses. One of them was discussed above (see Section 4-2). The process in a gas of constant mass which proceeds in a constant volume is termed *isochoric* (from the Greek *choros* for clear space). The plots of this process are termed *isochores* (see Fig. 4.2).

Note that the combined gas law and formulae (5.3), (5.7) and (5.8) are applicable to any isoprocess with account taken of the fact that in each case one of the properties remains constant. In an isochoric process the parameter remaining constant is  $V$ ; therefore after  $V$  is cancelled out formula (5.3) assumes the form

$$p_1/T_1 = p_2/T_2, \quad \text{or} \quad p_1/p_2 = T_1/T_2 \quad (5.15)$$

Hence, the isochoric process obeys the following law: for constant mass and volume the pressure of a gas is directly proportional to its absolute temperature. This also follows from the ideal-gas law, Eq. (5.7):

$$pV = \frac{m}{\mu} RT$$

Since  $V$ ,  $m$ ,  $\mu$  and  $R$  remain constant, it follows from (5.7) that  $p$  is proportional to  $T$ . Note that the law may also be formulated in the way it was done in Section 4-2.

## 5-7 Isobaric Process

The process in a gas of constant mass which proceeds at a constant pressure is termed *isobaric* (from the Greek *baros* for weight). This process was studied by the French physicist J. L. Gay-Lussac (1778-1850) in 1802.

Since in an isobaric process  $p$  remains constant, Eq. (5.3), after  $p$  is cancelled out, assumes the form

$$V_1/T_1 = V_2/T_2, \quad \text{or} \quad V_1/V_2 = T_1/T_2 \quad (5.16)$$

Equation (5.16) is the mathematical expression for the *Gay-Lussac law*: for a constant mass of gas at constant pressure its volume is directly proportional to its absolute temperature. The law is sometimes called the *Charles law*.

Figure 5.1 is a schematic diagram of Gay-Lussac's experiment. A bulb containing gas is placed into a tank holding

Fig. 5.1 Heating gas at constant pressure.

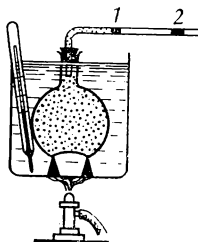
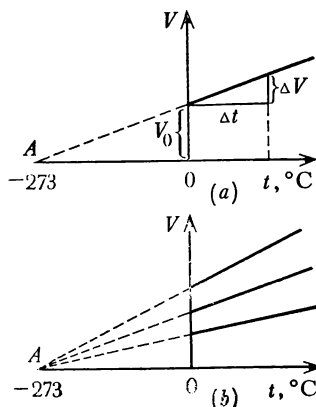


Fig. 5.2 (a) Plot of isobaric process for ideal gas; (b) isobars of different gases intersect the  $t$  axis at one point.



water and ice. A bent tube with the horizontal end open passes through the plug. The gas in the bulb is separated from the atmosphere by a small column of mercury in the tube. To measure the temperature of the gas a thermometer is employed, and the position of the column of mercury serves as the indication of the volume. For this reason the tube is provided with divisions corresponding to a definite internal volume of the tube (when the tube is being graduated account can also be taken of the thermal expansion of the bulb, but this is comparatively small and can be neglected).

First the volume of the gas at  $0^\circ\text{C}$ ,  $V_0$ , is determined from the position of the mercury column 1. Next the gas is heated (the mercury column moves to position 2) and the values of volume and temperature are recorded in the process. Finally a graph is plotted; this graph is termed *isobar*.

The isobar turns out to be a straight line (Fig. 5.2a) intersecting the  $t$  axis at point A corresponding to  $-273^\circ\text{C}$ . Should this experiment be repeated with various gases or with different masses of the same gas, all plots would be found to intersect at point A (Fig. 5.2b). This means that the expansion of a gas in an isobaric process is independent of its nature.

It follows from the similarity of triangles in Fig. 5.2a that

$$\frac{V_0}{OA} = \frac{\Delta V}{\Delta t}, \quad \text{or} \quad \frac{1}{OA} = \frac{\Delta V}{V_0 \Delta t}$$

Denoting  $1/OA = 1/(273^\circ\text{C})$  by  $\beta$ , we obtain

$$\Delta V = \beta V_0 \Delta t \quad (5.17)$$

Here  $\beta$  is the coefficient of volume expansion at constant pressure, or simply the *expansion coefficient*, since there is no linear coefficient to be confused with it (see Chapter 15).

Note that for gases the coefficients  $\gamma$  and  $\beta$  in Eqs. (4.3) and (5.17) are numerically equal, therefore one letter is used to denote them and only the coefficient  $\beta$  is used in calculations.

## 5-8 Isothermal Process

The process in a gas of constant mass which proceeds at a constant temperature is termed *isothermal*.

Isothermal processes in gases were studied by the Irish scientist Robert Boyle (1627-1691) and the French physicist Edmé Mariotte (c. 1620-1684) (Boyle published his results

in 1663 and Mariotte in 1676). The relation which they obtained from experiment follows directly from formula (5.3) after  $T$  is cancelled out:

$$p_1 V_1 = p_2 V_2, \quad \text{or} \quad \frac{p_1}{p_2} = \frac{V_2}{V_1} \quad (5.18)$$

Equation (5.18) is the mathematical expression for the *Boyle law*: for a constant mass of a gas at a constant temperature its pressure is inversely proportional to its volume. In other words, under the conditions specified the product of the volume of a gas by its pressure is a constant:

$$pV = \text{constant} \quad (5.19)$$

The relation (5.19) may be obtained either from (5.7) or from (5.8), since for constant  $T$  the right-hand sides of (5.7) and (5.8) are constants. The plot of the  $p$  versus  $V$  dependence for an isothermal process is a hyperbola and

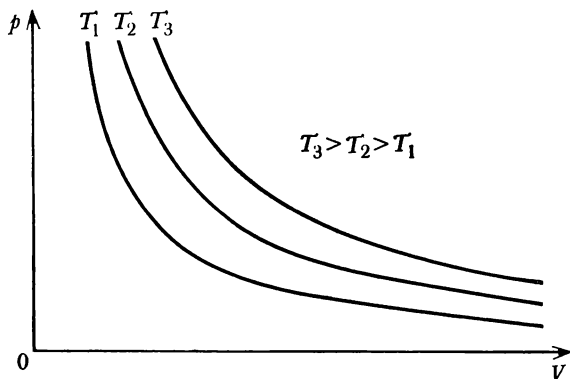


Fig. 5.3 Isothermal curves for same masses of ideal gas at different temperatures.

is termed an *isotherm*. Figure 5.3 shows three isotherms for the same mass of a gas but for different temperatures  $T$ .

Note, in addition, that a direct result of Eq. (5.9) is that in an isothermal process the density of a gas changes in direct proportion to its pressure:

$$\rho_1 / \rho_2 = p_1 / p_2 \quad (5.20)$$

Experimental verification of the Boyle law presents no problems. (Ponder on the question of how this can be done.)

## 5-9 Internal Energy of Ideal Gas

As was remarked in Section 4-1, there are no forces of interaction between the molecules in an ideal gas. This means that an ideal gas has no molecular potential energy.

Besides, the atoms of an ideal gas are particles, that is, they are devoid of internal structure and, consequently, of energy due to the motion and interaction of particles making them up.

Hence, internal energy of an ideal gas is just the sum of the values of kinetic energy of random motion of all its molecules.

$$E = \sum K_i$$

Since there is no such thing as rotation of a particle, the molecules of monatomic gases take part only in translational motion. Since the expression for the average value of the energy of translational motion of molecules is

$$\bar{K}_{\text{trans}} = \frac{3}{2} k_B T$$

(see Eq. (4.8)), it follows that the energy of a mole of monatomic ideal gas is

$$E_{\text{mole}} = \frac{3}{2} N_A k_B T$$

where  $N_A$  is the Avogadro number. Since  $N_A k_B = R$ , we obtain

$$E_{\text{mole}} = \frac{3}{2} RT \quad (5.21)$$

For an arbitrary mass of monatomic ideal gas we have

$$E_1 = \frac{3}{2} \frac{m}{\mu} RT \quad (5.22)$$

If the molecule is made up of two rigidly joined atoms (diatomic gas), its molecules in the course of random motion also take part in rotational motion about two mutually perpendicular axes. Because of that at the same temperature the internal energy of a diatomic gas is greater than that of a monatomic gas:

$$E_2 = \frac{5}{2} \frac{m}{\mu} RT \quad (5.23)$$

Finally the internal energy of a multiatomic gas (whose molecules are made up of three or more atoms) is twice that of the monatomic gas at the same temperature

$$E_m = 3 \frac{m}{\mu} RT \quad (5.24)$$

since the contribution of molecular rotation about three mutually perpendicular axes is the same as that of its translational motion in three mutually perpendicular directions.

Note that Eqs. (5.23) and (5.24) do not hold for real gases at high temperatures since then vibrations of atoms in the molecules are excited, which results in an increase in the internal energy of the gas. (Why does not this apply to Eq. (5.22)?)

### 5-10 Work Performed by Gas

Experience shows that a compressed gas can perform work in the course of its expansion. Instruments and devices whose operation is based on this property of gas are termed *pneumatic*. This is the principle of operation of the pneumatic drill, of mechanisms for opening and closing doors in public transport and of other devices.

Imagine a cylinder filled with gas and containing a mobile piston (Fig. 5.4). As long as the pressure of the gas inside

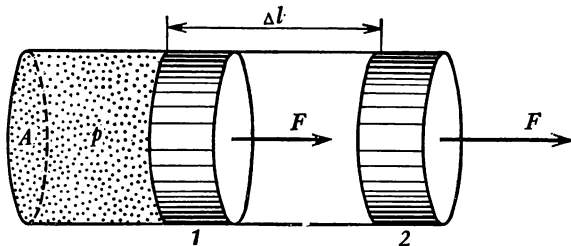


Fig. 5.4 In the course of isobaric expansion, gas performs work in displacing piston from position 1 to position 2.

the cylinder and that of the atmosphere are equal, the piston is at rest. Let the temperature of the gas and of the surroundings be  $T_1$  and the pressure be  $p$ .

Let us now slowly heat the gas in the cylinder until it attains the temperature  $T_2$ . External pressure  $p$  remaining constant, the gas will expand at constant pressure with the piston travelling the distance  $\Delta l$  from position 1 to position 2 the gas performing work against external force. The force that performs the work is  $pA$ , where  $A$  is the cross section of the cylinder. The formula for work established in mechanics is

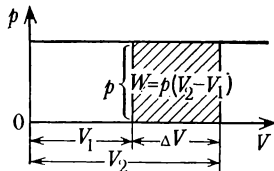
$$W = F\Delta l, \text{ or } W = pA\Delta l$$

Since  $A\Delta l$  is the increment of volume of the gas in the process of its heating from  $T_1$  to  $T_2$  at a constant pressure, we have

$$W = p\Delta V, \text{ or } W = p(V_2 - V_1) \quad (5.25)$$

It can easily be seen that the work performed in an isochoric process is always zero, since in this case there is no

Fig. 5.5 Isobaric process (shaded area is numerically equal to work performed by gas).



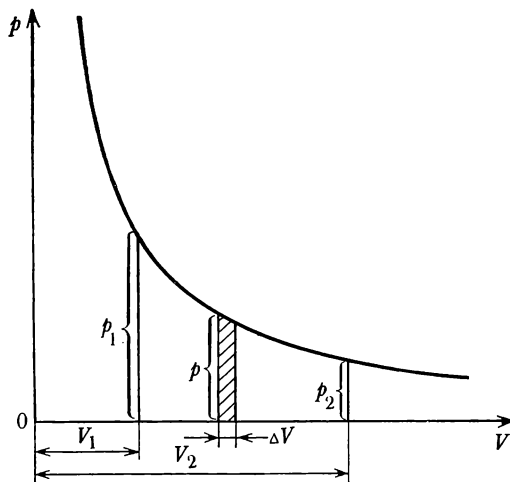
variation in volume. In general, it should be remembered that a gas performs work only in the process of changing its volume, that is, when  $\Delta V \neq 0$ . Now it is obvious that in an isobaric process the gas performs work (expressed by Eq. (5.25)) and that in an isochoric process it does not.

Note that when the gas is compressed ( $\Delta V$  is negative), positive work is performed by external forces, the work performed by the gas being in this case *negative*. In this case the internal energy of the gas should increase at the expense of the work done by external forces. When the gas expands ( $\Delta V$  is positive), the work done by the gas is *positive* and its energy should decrease by the amount of work performed,  $W$ . Naturally, the energy of the surroundings in this case increases by the same amount.

Let us now find how to determine the work of a gas using the  $p$  versus  $V$  plot for the process. In the case of an isobaric process the  $p$  versus  $V$  plot is a straight line parallel to the  $V$  axis (Fig. 5.5), since in this case  $p$  is constant. It follows from Fig. 5.5 that the work is numerically equal to the shaded area.

Now let us turn to the isothermal process. Figure 5.6 shows an isotherm of an ideal gas. The gas performs work

Fig. 5.6 Isothermal process (work performed by gas,  $\Delta W$ , is numerically equal to shaded area).



in this process, since  $\Delta V$  is not zero. Equation (5.25) cannot be applied in this case since it is valid for  $p = \text{constant}$  and  $p$  changes in the isothermic process. One may, however, take such a small increment of the volume,  $\Delta V$ , that the variation of the pressure can be neglected. In this case

the work  $\Delta W$ , can be computed with the aid of the formula

$$\Delta W = p \Delta V$$

In Fig. 5.6 this is illustrated by a shaded area.

If we divide the interval  $(V_2 - V_1)$  into intervals  $\Delta V$  so small that the formula  $\Delta W = p \Delta V$  may be used to compute work in each of them, we shall find the total work performed by the gas as the sum of the elementary work increments,  $\Delta W$ . This will mean that the work of the gas will be equal to the sum of areas like the shaded area in Fig. 5.6. Therefore in the isothermal process the work of a gas is expressed by the area bounded by two ordinates  $p_1$  and  $p_2$ , the  $V$  axis and the  $p$  versus  $V$  plot.

It may be proved rigorously that work of a gas in any process is expressed by the area bounded by two ordinates, by the abscissa and the plot of this process in the  $V$ - $p$  coordinates.

Let us now find the physical meaning of the universal gas constant,  $R$ . Applying Eq. (5.25) to a mole of an ideal gas, we obtain

$$W_{\text{mole}} = p \Delta V_{\text{mole}} \quad (5.26)$$

But with the aid of the ideal-gas law (5.8) we may write for two states of a mole of gas:

$$p V_{2 \text{ mole}} = R T_2 \quad \text{and} \quad p V_{1 \text{ mole}} = R T_1$$

Whence

$$p (V_{2 \text{ mole}} - V_{1 \text{ mole}}) = R (T_2 - T_1), \quad \text{or} \quad p \Delta V_{\text{mole}} = R \Delta T$$

Substituting into expression (5.26), we obtain

$$W_{\text{mole}} = R \Delta T, \quad \text{or} \quad R = W_{\text{mole}} / \Delta T \quad (5.27)$$

It follows from (5.27) that the universal gas constant is numerically equal to isobaric work performed by a mole of ideal gas heated one kelvin.

It follows from the relation  $k_B = R/N_A$  that the Boltzmann constant is the average work of a molecule of ideal gas at constant pressure heated one kelvin.



# 6

## Internal Energy

### 6-1 Internal Energy and the Surroundings

The initial concept of the internal energy of a body was outlined in Section 2-6. Let us return to this very important concept and refine it. The internal energy of a body is understood to be the sum of the kinetic and potential energies of all the particles making up the body and the energy of the nuclei of its atoms. At this stage the question is: What energy is not a part of internal energy of a body?

Let us use the Earth as an example of such a body. It is an established fact that the Earth is attracted to the Sun and moves in an orbit around it, that is, possesses both potential and kinetic energy with respect to the Sun. This energy is not a part of the internal energy of the Earth since it owes its existence to another body outside the Earth. Hence, the internal energy of a body does not include kinetic and potential energy of that body with respect to all external bodies. In the same way the internal energy of a system of bodies does not include kinetic and potential energy of this system with respect to bodies that are not a part of this system.

Of no less importance is the problem of calculating the internal energy of a body or of a system of bodies. Actually, an exact formula exists only for an ideal gas (see Section 5-9). In all other cases we are unable to calculate the internal energy. However, of major importance for practical purposes is not the internal energy itself but its variation, that is, the difference of its values at the beginning and at the end of some process. To find this difference one need not know the numerical value of the total internal energy. This is true not only of internal energy. Another example is the temperature difference, which is the same in the Celsius and the thermodynamic scales, that is, is independent of the numerical value of the temperature itself, provided the units of measurement do not change.

Since molecular physics deals only with such phenomena in which the molecules are not subject to change, it may be assumed that only molecular kinetic and potential energies vary in those phenomena. This substantially simplifies most calculations.

On account of the energy conservation law we can say that the variation of internal energy of a body is always due to its interaction with other bodies and with the surroundings.

In some cases the variation of the internal energy of a body is found from the (known) amounts of the energy lost or gained by the body as the result of its interaction with the surroundings. In other cases, on the contrary, the amounts of energy gained by the surroundings and by other interacting bodies is found from the variations of internal energy of a body. One of the most important forms of energy transfer between bodies is heat exchange.

## 6-2 Heat Exchange

A hot kettle standing on the table cools after some time. At daytime the Earth's surface is noticeably heated by sunshine. The metal handle of a frying pan standing on a stove becomes very hot. All those are examples of heat exchange. Exchange of internal energy between bodies and environment, or between different parts of a body in cases when no mechanical work is performed is termed *heat exchange*.

Energy transfer in the course of heat exchange is the result of numerous individual acts of interaction between molecules; in other words, it is the result of numerous microprocesses. For instance, the cooling of hot water is explained by the energy transfer resulting from collisions between the molecules of water and those of air surrounding the water, the heating of air and the cooling of water being due to the fact that in the majority of such collisions energy is lost by the water molecules and gained by the air molecules. However, in some collisions the air molecules may lose energy and the water molecules gain energy, since the energy of certain molecules may be substantially different from the average value (see Section 3-4). The frequency of such events will rise as the temperatures of water and air level out. Then, when their temperatures become equal, the events in which the air molecules gain or lose energy in collisions with the water molecules will be equally probable and there will be no energy transfer from water to air or back.

Thus, most acts of molecular interaction in the process of heat exchange result in the energy being transferred from the body with the greater temperature to the body with the lower temperature, which, in turn, leads to the equalizing of the temperatures of those bodies. Note that this assertion expresses a statistical law and will not prove true if only a small number of molecules takes part in the interactions.

In accordance with historical tradition the variation of internal energy of a body is often referred to as the *quantity of heat*  $Q$ , gained or lost. If a body's internal energy rose in the process of heat exchange by  $\Delta U$ , the body is said to have gained a quantity of heat  $Q$ , and if the internal energy has decreased by  $\Delta U$ , the body is said to have lost heat  $Q$ . Hence,  $Q$  is the numerical value of the energy transmitted or gained by the body in the process of heat exchange. The unit for measuring heat in the SI system is the *joule* (J). Note that formerly the units for measuring  $Q$  were the calorie (cal) and the kilocalorie (kcal) (see Section 8-1).

Finally, let us stress once again that the quantity of heat,  $Q$ , is a measure of the variation of internal energies of bodies taking part in a process of heat exchange and that it depends essentially on the type of the process. This means that one can speak about quantities of heat only in connection with a specific process. The concept of heat contained in a body that remains in some definite state is absolutely meaningless. In that case we can speak only of the internal energy of this body.

### 6-3 Types of Heat Exchange

In nature heat exchange is realized in the forms of heat conduction, radiation (absorption and emission of radiation) and convection.

*Heat conduction* is the term for the transfer of internal energy from some parts of a substance to other parts, resulting from the random motion of molecules and other particles making up the substance.

The mechanism of heat conduction was actually explained in the preceding section. Here is an additional example. When the end of a metal rod is heated, the speed of its molecules rises, that is, the internal energy of this end of the rod increases. Since the molecules of its opposite end move at lower speeds, random motion of the atoms and the electrons inside the rod brings about the transfer of internal energy to the colder end. This raises its temperature.

Among various kinds of substances metals have the greatest heat conductions. They owe this to the free electrons they contain. Note in addition that the heat conduction of a substance in the solid state exceeds that of its liquid state and the latter, in turn, exceeds that of the gaseous state. (Explain why.)

Let us find how the temperature of all objects contained in an unheated room levels out. It cannot be explained by

heat conduction alone since heat conduction of air is small (we recall that double panes are used for better heat insulation in winter) and the temperatures of the objects level out quite quickly. It turns out that a major role in the process belongs to electromagnetic radiation. The motion of electric charges in the atoms and molecules of all substances creates electromagnetic radiation whose intensity rises sharply with temperature. The radiation process takes place at the expense of the internal energy of the radiating body, which decreases in the process. When a body absorbs such radiation, it increases its internal energy.

Thus all the objects inside the room at the same time emit and absorb radiation. In the process the objects with higher temperatures lose more energy than they gain and cool down and the objects with lower energy gain more energy than they lose and are heated. Such energy transfer exists at all temperatures (it may cease only at absolute zero) and makes the temperature of all the objects the same. This will be discussed in more detail in Chapter 37.

Heat exchange by means of radiation takes place even if there is no substance between the bodies. For instance, the Earth receives energy from the Sun by means of radiation travelling through space. Radiation of bodies due entirely to their temperature is termed *thermal radiation*.

Let us now discuss convection. To demonstrate poor heat conduction of water the vessel containing it is usually heated from above. In this case water at the top may be boiling and still stay cool at the bottom. However, should we heat the water from the bottom it would be heated uniformly throughout the vessel's volume. The explanation is that water expands on heating and its density drops. If the heated water is at the bottom, the upper denser layers will sink due to the force of gravity and will take the place of warm water. Such mixing will continue until all water begins to boil. Heat exchange due to the mixing of nonuniformly heated layers of a fluid acted upon by the force of gravity is termed *convection*.

It may easily be guessed that there should be no convection in a spacecraft that is in a state of weightlessness. (Ponder on the question why the freezer in a refrigerator is mounted above and not below.)

It may seem at this point that convection cannot be classified as heat exchange since it is due to the work of the gravitational force. However, the increase in the internal energy of a liquid or gas in the process of convection is due entirely to the heat supplied from outside, the effect of the

force of gravity being only to accelerate the uniform heating of the liquid or gas. In the process of convection the force of gravity does not contribute to the internal energy of a liquid or gas. This is the reason why convection is classified as heat exchange.

#### **6-4 Changing Internal Energy by Means of Work**

Let us discuss several examples to the effect that internal energy can change not only in the course of heat exchange but also when mechanical work is performed.

When wood is sawn by a handsaw, the saw gets hotter. When some mechanical part is being drilled, both the drill and the part become quite hot. The cutting tool of a lathe becomes hot in the process of machining. Innumerable examples of this sort may be cited. They all show that every time work is performed to overcome friction or to break up some material various objects are heated, that is, their internal energy is raised in the same way as when they gain heat. Therefore, in the above examples, work is said to be transformed into heat. In those cases mechanical energy of bodies (which is not a part of their internal energy) is transformed into their internal energy, which results in the bodies being heated.

In the eighteenth century the caloric theory was used to explain thermal phenomena. Scientists thought that there was a special weightless kind of substance, caloric, whose amount in nature remains constant. It was postulated that when a body is cooled, the caloric flows from the body to the surroundings, and when the body is heated, the caloric enters it from other bodies.

The great Russian scientist M. V. Lomonosov (1711-1765) was one of the first to speak out against the caloric theory. He explained thermal phenomena as being due to the motion of invisible particles of the bodies. In 1798 Benjamin Thompson (1753-1814), an American who later became Count Rumford of Bavaria, demonstrated that unlimited heat may be obtained in the course of boring cannon (at the expense of mechanical work). Rumford's results proved that the caloric theory was wrong. Of no less importance for the refutation of the caloric theory were the results obtained by the French engineer Sadi Carnot (1796-1832), who evolved the theory of heat engines.

### 6-5 Relation of Internal Energy to State of Matter

It was established in the course of studies of properties of matter that the same substance may be encountered in the solid, liquid and gaseous states, for instance, ice, water and vapour. The general term for such states is *states of aggregation*, or simply *states of matter*. It was established that the state of aggregation is connected with the internal energy of a unit of mass of the substance. In the liquid state this energy is greater than in the solid, and in the gaseous state it is greater than in the liquid.

Therefore the transition from one state of aggregation of the substance into another should involve noticeable changes in the internal energy of this substance. Let us discuss from the standpoint of the kinetic theory what happens to a substance when its internal energy is being gradually raised.

Consider the potential curve for two molecules (see Section 2-5) which determines the potential energy of those molecules as a function of the distance  $r$  between them (Fig. 2.3b). In the absence of kinetic energy the molecules would occupy positions at a distance  $r_0$  corresponding to their stable equilibrium, since in this case the resultant of the forces of molecular interaction is zero (Fig. 2.2). Actually the molecules are in motion and the distance between them changes continuously. When the kinetic energy of translational motion of a molecule,  $\bar{K}_{\text{trans}}$ , is small, the molecule vibrates. In the process its kinetic energy is transformed into potential and vice versa. Since the plot of the potential energy to the left of  $r_0$  rises steeply and is flatter to the right of  $r_0$ , the average distance between the vibrating molecules,  $r_1$ , exceeds  $r_0$ .

When a substance is heated, the average kinetic energy of translational motion of its molecules,  $\bar{K}_{\text{trans}}$ , increases together with the average distance,  $\bar{r}$ , between the molecules. The temperature corresponding to  $r_2$  in Fig. 2.3b is  $T_2$  and that to  $r_1$  is  $T_1$  ( $T_2 > T_1$ ). Hence, the substance expands upon heating and contracts upon cooling.

As long as the average kinetic energy remains sufficiently small (usually much smaller) as compared to  $U_{\text{min}}$ , the substance remains solid. When in the process of heating (of raising its internal energy)  $\bar{K}_{\text{trans}}$  rises above a certain level, the substance turns into a liquid. When the internal energy and  $\bar{K}_{\text{trans}}$  increase still further, the substance turns

into a gas, since the forces of molecular interaction are no longer able to hold together the more energetic molecules.

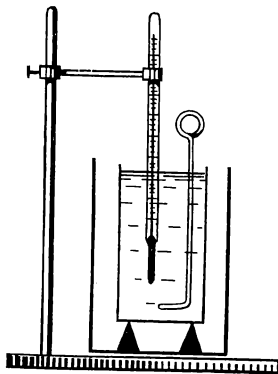
Roughly speaking, if  $\bar{K}_{\text{trans}} \ll U_{\text{min}}$ , the substance is a solid; if  $\bar{K}_{\text{trans}} \approx U_{\text{min}}$ , it is a liquid; and if  $\bar{K}_{\text{trans}} \gg U_{\text{min}}$ , it is a gas.

At this point another question arises: Why at the same temperature some substances are solid, others liquid and still others gaseous? The explanation is that although the shape of the potential curve is similar for molecules of all substances, the depth of the potential well,  $U_{\text{min}}$ , and  $r_0$  depend on the nature of the substance. Therefore at a definite temperature for some substances  $\bar{K}_{\text{trans}} \ll U_{\text{min}}$ , for others  $\bar{K}_{\text{trans}} \approx U_{\text{min}}$  and still for others  $\bar{K}_{\text{trans}} \gg U_{\text{min}}$ . This explains why at a definite temperature some substances will be in the solid state, some in the liquid and others in the gaseous. Of course, the explanation given here is purely qualitative.

## 7 Quantity of Heat

### 7-1 The Measurement of Heat

Fig. 7.1 Calorimeter.



Let us now find how the variation of internal energy in the course of heat exchange can be calculated. To make those calculations as accurate as possible the unaccountable heat losses in the process of heat exchange should be reduced to the minimum. Therefore in scientific research the heat exchange is effected in a calorimeter (Fig. 7.1), which makes possible a sufficiently accurate determination of the heat  $Q$  lost or gained by a body in the course of heat exchange.

The calorimeter consists of two vessels: the external and the internal. The internal vessel is manufactured from a good heat conductor (brass, copper) since its temperature should be the same as that of the liquid contained in it. The purpose of the external vessel is to prevent heat losses by convection and radiation from the internal vessel. To this end it is usually painted white or manufactured from bright tin. To minimize heat losses of the internal vessel by heat conductivity it is mounted on wooden supports (wood is a poor conductor of heat). A mixer made of the same material as the internal vessel and a thermometer are immersed in it.

The heat exchange experiment is effected in the following manner. The masses of the calorimeter's internal vessel and

of the mixer are determined with the aid of scales. The mass of the liquid contained in the calorimeter, for instance, water, is measured in the same way. After that the mass of the body chosen for the heat exchange experiment is measured, and the body is heated to a definite temperature and immersed into the liquid contained in the calorimeter whose temperature was previously measured. In this way the amount of heat lost by the body in the course of heat exchange can easily be measured.

## 7-2 Changing Internal Energy by Heating or Cooling

Consider the heat exchange the sole result of which is the heating or the cooling of bodies. Using a calorimeter one can easily establish that under those conditions the variation of internal energy of a body is proportional to its mass  $m$  and to the temperature variation  $\Delta T$ :

$$\Delta E = cm \Delta T \quad (7.1)$$

where  $c$  is a proportionality factor. Since the measure for the variation of internal energy in the course of heat exchange is the heat  $Q$ , we have

$$Q = cm \Delta T \quad (7.2)$$

The initial temperature of the body is usually denoted by  $T_1$  and the final by  $T_2$ . Then, if the body is heated,  $\Delta T = T_2 - T_1$ , and if it is cooled, it is reasonable to assume that  $\Delta T = T_1 - T_2$ .

In order to find whether  $Q$  in Eq. (7.2) depends on the nature of substance, one may make the following experiment. Take several cylinders or bars of equal mass but made of different material, heat them to an equal temperature and after that place them quickly on a paraffin block. The amounts of paraffin observed to have melted under each cylinder in the process of its cooling down to room temperature will be different. This means that  $Q$  depends on the nature of the substance. Experiments show that  $Q$  depends, besides, on external conditions, for instance, on the initial temperature of the body and on the state of aggregation of its substance. Those factors are taken into account by  $c$  of Eqs. 7.1) and (7.2)

The quantity that characterizes the dependence of the variation of internal energy of a body being heated or cooled on its substance and on the external conditions is termed *specific heat* of the substance. Specific heat is the quantity



of heat needed to heat a unit of mass of the substance per degree:

$$c = \frac{Q}{m \Delta T} \quad (7.2a)$$

Let us derive the unit for measuring specific heat

$$c = \frac{1 \text{ J}}{1 \text{ kg} \cdot 1 \text{ K}} = 1 \frac{\text{J}}{\text{kg} \cdot \text{K}}$$

In the SI system the accepted unit of specific heat is the specific heat of such a substance that requires energy of 1 J to raise the temperature of 1 kg of its mass by 1 K. For small temperature variations the specific heat may be presumed to be a constant. For solving problems its values should be taken from appropriate tables.

One should keep in mind that in determining the amounts of heat needed to heat a body or lost by it in the process of cooling use is sometimes made of the *thermal capacity* of a body,  $C$ , which is the heat needed to raise the temperature of a body by one degree. Therefore

$$Q = C \Delta T \quad (7.3)$$

The use of the thermal capacity of a body is especially convenient when different parts of the body are made up of different materials. The unit of thermal capacity in the SI system is 1 J/K. (Demonstrate it with the aid of Eq. (7.3).)

Note in addition that the specific heat of a gas depends on the nature of the process in which it is heated. For instance, the specific heat of a gas at constant pressure,  $c_p$ , exceeds its specific heat at constant volume,  $c_v$ , since in the first case not only must the temperature of the gas be raised but it must in addition perform work of expansion against external pressure (see Section 5-10). In the second case the entire heat transmitted to the gas is spent on increasing its internal energy.

### 7-3 Heat of Combustion

The internal energy of a body can partially be liberated when it takes part in chemical reactions. A considerable quantity of heat is liberated in the combustion process. Substances used to produce heat are called *fuels*. Energy liberated in the process of fuel combustion is widely used in industry, transport and everyday life. There are solid, liquid and gaseous fuels.

Experiments show that the heat liberated in the combustion of a specific kind of fuel,  $Q_{\text{lib}}$ , is directly proportional to the mass of the fuel consumed,  $m$ :

$$Q_{\text{lib}} = qm \quad (7.4)$$

There is a qualitative difference between various kinds of fuel primarily in the value of  $Q_{\text{lib}}$  and, consequently, in the value of the proportionality factor  $q$  in (7.4).

The factor  $q$  characterizing the dependence of heat liberated in the process of fuel combustion on the kind of fuel is termed *specific heat of combustion*. Specific heat of combustion is the quantity of heat liberated as the result of complete combustion of a unit mass of the fuel:

$$q = Q_{\text{lib}}/m \quad (7.4a)$$

Let us find the unit for measuring  $q$  in the SI system:

$$q = 1 \text{ J/1 kg} = 1 \text{ J/kg}$$

Equation (7.4) may be conveniently used for calculating heat liberated in the process of combustion of solid and liquid fuels. Presently gaseous fuel is being more and more used and for it Eq. (7.4) is not quite suitable since it is more convenient to express the consumption of gaseous fuel in cubic meters. However, the pressure in a gas pipeline is above the atmospheric and the volume of a gas depends on external conditions. Therefore gas meters are designed to show the volume of gas consumed in cubic meters under standard conditions,  $V_0$ . Since the heat  $Q$  liberated in the combustion of a gas is directly proportional to  $V_0$ , it follows that

$$Q_{\text{lib}} = \kappa V_0 \quad (7.5)$$

Here  $\kappa$  (the Greek *kappa*) is the *specific heat of combustion of gaseous fuel* and depends on the nature of the gas. The term specific heat of combustion of gaseous fuel means the amount of heat liberated in the process of complete combustion of a unit volume of the gas under standard conditions:

$$\kappa = Q_{\text{lib}}/V_0$$

In the SI system  $\kappa$  is measured in  $\text{J/m}^3$ .

The requirements for gas in the world are becoming greater and greater. Every year new gas fields are being put into operation and new pipelines are being constructed. Natural gas is also an important raw material for chemical synthesis.

Note that fuel requirements are often expressed in tons of *conventional fuel* whose specific heat of combustion was, by agreement, fixed at  $29.3 \times 10^6 \text{ J/kg}$  (7000 kcal/kg).

Fuel is burnt in furnaces, boilers, nozzles, gas stoves, etc. the general name for which is *heaters*. The design of the apparatus for burning fuel is determined mainly by the fuel being used and by the purpose for which the heat is produced. It is not possible, however, to use all of the heat produced by the heater, since some of the heat is carried away by the combustion products and is dispersed in the surrounding medium.

The quantity  $\eta$  characterizing the effectiveness of the fuel-burning apparatus is termed the *efficiency* of this apparatus. The efficiency of a heater shows what part of the heat liberated in the process of fuel combustion,  $Q_{\text{lib}}$ , constitutes the useful heat,  $Q_{\text{use}}$ :

$$\eta = Q_{\text{use}}/Q_{\text{lib}} \quad (7.6)$$

or

$$\eta = (Q_{\text{use}}/Q_{\text{lib}}) \times 100\% \quad (7.6a)$$

Note that  $Q_{\text{lib}}$  is always found from Eqs. (7.4) or (7.5), whereas  $Q_{\text{use}}$  depends on the type of the fuel-burning apparatus and may be expressed by various formulae.

#### 7-4 The Law of Heat Exchange

First let us recall that in the absence of mechanical work the variation of the internal energy of a body is measured by the heat  $Q$ . Since in the case of a pure heat exchange there are no other mechanisms for changing the internal energy, we can say that in this case some bodies will lose as much heat as the others will gain.

This serves as a basis for the law of heat exchange which is used whenever the quantity of heat must be calculated. When writing out this law, one should keep in mind that in the course of heat exchange the sum of the quantities of heat lost by all bodies whose internal energies decrease in the process is equal to the sum of the quantities of heat gained by other bodies whose internal energies increase in the process:

$$\sum Q_{\text{lost}} = \sum Q_{\text{gained}} \quad (7.7)$$

Heat exchange continues until the temperatures of all the bodies become equal. The common temperature after the termination of heat exchange is denoted by  $\Theta$  (the Greek *theta*).

As an example, let us write down the law of heat exchange used for determining specific heat of a substance with the

aid of a calorimeter. We can approximately assume that in this case three bodies take part in the heat exchange: the calorimeter, the liquid and the body whose specific heat is being determined. This body is initially heated to a definite temperature  $T_2$  and is then immersed into the liquid with a temperature  $T_1$  in the calorimeter. After some time a common temperature  $\Theta$  is established in the calorimeter.

In that case it may be asserted that the body gave up heat  $Q_{\text{body}}$  and that the calorimeter and the liquid gained quantities of heat equal to  $Q_{\text{cal}}$  and  $Q_{\text{liq}}$ , respectively. Then

$$Q_{\text{body}} = Q_{\text{cal}} + Q_{\text{liq}}$$

Since  $Q_{\text{body}} = c_{\text{body}} m_{\text{body}} (T_2 - \Theta)$ ,  $Q_{\text{cal}} = c_{\text{cal}} m_{\text{cal}} \times (\Theta - T_1)$  and  $Q_{\text{liq}} = c_{\text{liq}} m_{\text{liq}} (\Theta - T_1)$ , we have

$$c_{\text{body}} m_{\text{body}} (T_2 - \Theta) = c_{\text{cal}} m_{\text{cal}} (\Theta - T_1) + c_{\text{liq}} m_{\text{liq}} (\Theta - T_1)$$

or

$$c_{\text{body}} = \frac{(\Theta - T_1) (c_{\text{cal}} m_{\text{cal}} + c_{\text{liq}} m_{\text{liq}})}{m_{\text{body}} (T_2 - \Theta)}$$

Substituting the numerical values obtained in the experiment into the right-hand side of the last formula, we can calculate the sought specific heat.

## The Law of Conservation of Energy. The First Law of Thermodynamics

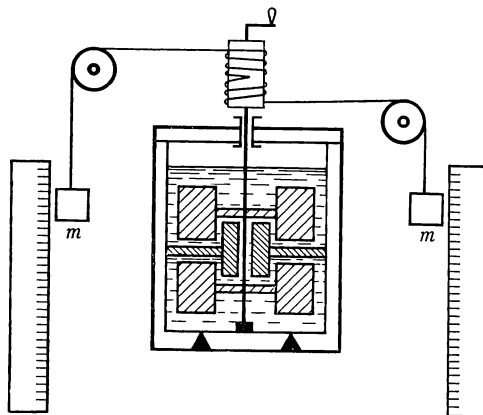
## 8

### 8-1 Mechanical Equivalent of Heat

It was stated above that unlimited heat may be produced with the aid of mechanical work. At this point the question is: Is there a definite quantitative relation between mechanical work and heat? In other words, do equal amounts of work always produce the same amounts of heat? To obtain an answer to this question the British physicist James P. Joule (1818-1889) carried out a series of experiments. And the answer was positive. His first experiment dates back to 1843.

Figure 8.1 shows the setting of one of Joule's experiments. The apparatus consisted of a calorimeter filled with mercury. An axle passed through the calorimeter with a shaft and a handle on its end. A thread was wound around the shaft and equal weights of mass  $m$  were attached to its ends. Rulers

Fig. 8.1 Joule's experiment.



to measure the displacement of the weights were arranged alongside each weight. To increase friction a paddle system was inserted into the calorimeter, and this system was connected to the axle.

Prior to the experiment the weights were raised to their top positions by turning the handle and the temperature of the mercury was recorded. Next the handle was let free and the weights moving downwards caused the rotation of the paddles inside the calorimeter. Because of substantial friction between the paddles and the mercury and between the layers of mercury moving at different speeds (hydraulic friction) heat was liberated in the calorimeter at the expense of work performed by the weights along path  $h$ . Since friction in the parts of the apparatus outside the calorimeter was negligible, it can be assumed that the increase in the internal energy of the calorimeter (including the liquid and all the parts inside it) was equal to the decrease in the mechanical (potential) energy of the weights in the process of their motion.

The energy conservation law was not definitely established at that time, but Joule assumed that the heat  $Q$  liberated in the calorimeter is equal to the work of the weights,  $2 mgh$ . His experiments proved that the quantity of heat liberated by friction is directly proportional to the work performed.

According to modern data, to heat 1 kg of water from 292.5 K to 293.5 K, that is, by 1 K, we need 4186 J. This means that the specific heat of water in these conditions is

$$c_{\text{water}} = 4186 \text{ J}/(\text{kg} \cdot \text{K}) \approx 4200 \text{ J}/(\text{kg} \cdot \text{K})$$

Note that before the experiments of Joule only relative specific heats could be obtained from the heat exchange measurements, that is one could only find the ratio of specific heats of two substances, and the specific heat of water was chosen as the unit. The amount of heat required to heat 1 kg of water by 1 °C was termed *kilocalorie* (kcal). The expression for the specific heat of water was

$$c_{\text{water}} = 1 \text{ kcal}/(\text{kg} \cdot ^\circ\text{C}) = 1 \text{ cal}/(\text{g} \cdot ^\circ\text{C})$$

From Joule's experiments we have

$$1 \text{ kcal}/(\text{kg} \cdot ^\circ\text{C}) = 4186 \text{ J}/(\text{kg} \cdot \text{K})$$

This expression gives the relation between a kilocalorie and a joule:

$$1 \text{ kcal} = 4186 \text{ J} \approx 4200 \text{ J} = 4.2 \text{ kJ}$$

and

$$1 \text{ cal} \approx 4.2 \text{ J}$$

## 8-2 Conservation of Energy in Mechanics

Processes not accompanied by transformation of mechanical motion into other forms of motion we call *mechanical*. A system in which the forces acting between bodies depend only on the distances between them is called *conservative*. Such a system is an idealized system since it does not include the forces of friction and other resistance forces causing the dissipation of mechanical energy, that is, a transformation into other forms of energy.

In a conservative system the only possible transformation is that of the potential energy into the kinetic and back. The work of the forces acting in a conservative system is independent of the path and is only the function of the initial and the final positions of the body. The force of gravity is an example of such forces. It follows, then, that the work of a force in a conservative system along a closed path is zero. For such a system the following formulation of the *energy conservation law* is valid: in a closed conservative system the sum of the kinetic and potential energies of all the bodies making up the system is a constant. If this sum

is denoted by  $E$ , then in the absence of external influences

$$E = \text{constant} \quad (8.1)$$

For instance, the sum of kinetic and potential energies of a freely falling body is constant.

As we know, the sole measure of transfer of mechanical energy from one body to another is the work,  $W$ . Therefore, if the mechanical energy of a conservative system in some state is  $E_1$  and if subsequently external forces perform work  $W$  on it, the energy of the system increases by  $W$  and becomes  $E_2$  in its new state. Hence, in this case

$$E_2 - E_1 = W \quad (8.2)$$

Note, in addition, that in a wider sense the mechanical work in all natural phenomena serves as the sole measure of energy of mechanical motion transmitted (transformed) into other forms of motion and back.

### 8-3 The Law of Conservation of Energy

It was explained in Section 8-2 that mechanical energy is conserved only if there is no friction or other resistance forces. Action of the friction forces reduces mechanical energy. Indeed, a car gradually loses its kinetic energy and stops after its engine has been turned off; sledges gradually lose speed and stop at the bottom of the hill. It may easily be guessed that the apparent complete disappearance of the energy in all those cases is fiction. Detailed studies demonstrated that in all such cases some heat is produced, that is, bodies taking part in friction are heated and thereby increase their internal energy. Hence, friction and indeed any form of resistance is accompanied by the transformation of mechanical energy into internal energy.

Joule's experiments proved that  $W$  and  $Q$  are directly proportional (or equal if measured in the same units, say, in joules). Therefore the decrease in mechanical energy of the bodies in the presence of friction is exactly equal to the increase in internal energy of all the bodies taking part in such a process.

This means that the sum of the mechanical and internal energies of all the bodies making up a closed system is a constant. In other words, the total variation of the mechanical and internal energies of all the bodies in any process is zero if the variation is determined from the work performed and the heat transmitted.

Studies of natural phenomena have demonstrated that the internal energy of a body changes only as the result of work performed or energy exchanged with other bodies. In processes that we are dealing with now, heat exchange is the only form of energy transfer between the bodies. We can then say that the heat  $Q$  exchanged in the process and the work  $W$  done by the body uniquely determine the change in its internal energy.

The German physicist and physician Julius R. von Mayer (1814-1878) in 1842 drew attention to the fact that all forms of motion are mutually convertible. He then tried to apply the energy conservation principle to all natural phenomena. However, the man who placed this principle on a scientific footing was the German scientist Hermann von Helmholtz (1815-1888) in 1847.

Let us now formulate the law of conservation of energy: the energy of a closed system never vanishes and is never created out of nothing; in all phenomena taking place inside the system it merely changes its forms or is transmitted from one body to another, its total amount remaining constant. The discovery of this law was prompted by the idea of the perpetual motion machine, the *perpetuum mobile* (from the Latin for perpetual motion). Such a machine which would work on a closed-cycle principle and give more heat to surrounding bodies in one cycle than it would obtain from them is the *perpetual motion machine of the first kind*. It could serve as an inexhaustible source of energy and could work for any length of time without fuel.

Continuing failures of the inventors of such machines convinced scientists that its principle contravenes the laws of nature, the energy conservation law, to be exact. Therefore the energy conservation law can be expressed also in the following form: perpetual motion machine of the first kind is impossible.

The energy conservation law is a universal law of nature on which all branches of modern natural science are based. It is used to test new theories and assess the results of experiments. Should this law fail in some phenomena, we would have to review all other laws of natural science and to change our concept of the world.

#### 8-4 The First Law of Thermodynamics

There is a very important method for studying heat processes, the thermodynamic method. The essence of this method is as follows. In experiments, the numerical values of macro-



scopic quantities characterizing the process are measured. They are often termed thermodynamic properties (see Section 5-1). The results of the experiments are used to establish a relation between the parameters. Then a mathematical analysis of this relation is carried out on the basis of universal laws of nature whose validity is beyond doubt. The universal laws of nature used as a basis for such analysis are termed *laws of thermodynamics*.

For a successful analysis of the relations between the parameters the mathematical expressions for the laws of thermodynamics must be set in a specific form convenient for the analysis. Since all the conclusions of thermodynamics are based on experiments and on inviolable laws of nature, they never fail. It may easily be seen that the thermodynamic method enables the course of many phenomena to be predicted but fails to produce a model which could explain the physical meaning of the phenomena. This deficiency of the method is corrected by the kinetic theory, which explains many phenomena with the aid of such models. Hence, both major methods of research complement one another and together help us to gain a deep insight into the processes taking place.

The first universal law on which thermodynamics is based is the energy conservation law. This is called the *first law of thermodynamics* and is formulated as follows: heat  $Q$  supplied to a system is spent partially on increasing its internal energy by  $\Delta U$  and partially on work  $W$  performed by the system:

$$Q = \Delta U + W \quad (8.3)$$

Note that the system to which Eq. (8.3) may be applied can consist of a single body. The work  $W$  in (8.3) is numerically equal to the energy transferred to the system in the form of mechanical motion.

### 8-5 Some Applications of the First Law of Thermodynamics

Let us find what forms does formula (8.3) assume for the various isoprocesses taking place in ideal gas. We know already that in an isochoric process the work of a gas,  $W$ , is always zero. Therefore for this process formula (8.3) assumes the form

$$Q = \Delta U \quad (8.4)$$

This means that in an isochoric process all heat supplied to a gas is spent on increasing its internal energy (see Section 5-10).

For an isobaric process the formula expressing the first law of thermodynamics takes the form of (8.3):

$$Q = \Delta U + W$$

Indeed, in this case  $W = p\Delta V$  does not vanish since the volume of the gas does not remain constant in the process. The variation of the internal energy,  $\Delta U$ , which is proportional to the rise in temperature, is also nonzero since in this process the temperature is variable. Therefore in an isobaric process the heat supplied to a gas is spent partially on increasing its internal energy and partially on the work performed by the expanding gas.

In an isothermal process the temperature of the gas remains unchanged. Therefore, in accordance with Eqs. (5.22)-(5.24), the internal energy of the gas may be said to remain constant, i.e.  $\Delta U = 0$ . For this process Eq. (8.3) assumes the form

$$Q = W \quad (8.5)$$

This means that in an isothermal process all heat supplied to a gas is spent on work performed by it.

Analyzing formula (8.3), we see that yet another important process is possible with a gas, namely one in which  $Q = 0$ . This process is of great practical importance.

## 8-6 Adiabatic Process

A process taking place in a system without heat exchange with the surroundings is termed *adiabatic*. Since in this process  $Q = 0$ , Eq. (8.3) assumes the form

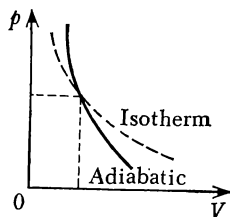
$$\Delta U + W = 0, \quad \text{or} \quad W = -\Delta U \quad (8.6)$$

This means that in an adiabatic process the gas performs work on external bodies only at the expense of its internal energy. Conversely, when in an adiabatic process external bodies perform work on the system, its internal energy increases.

If a gas performs work on the surroundings, its internal energy diminishes. Consequently, in this case it should cool down. (Ponder on the question whether a gas expanding into vacuum will cool down.) Adiabatic compression should obviously raise the temperature of the gas.

An equilibrium adiabatic process must, as all equilibrium processes, proceed at an extremely slow rate. It cannot be realized in practice, because there are no ideal heat insulators. However, the study of equilibrium adiabatic processes presents considerable interest since it enables important theoretical conclusions about the operation of heat engines

Fig. 8.2 Adiabatic process.



to be drawn. A plot of such a process in  $p$ - $V$  coordinates is shown in Fig. 8.2. The term for it is the *adiabatic*. For the sake of comparison, Fig. 8.2 depicts an isotherm for the same mass of ideal gas. (Why does the adiabatic run steeper than the isotherm?)

A nonequilibrium process in a gas under conditions of good thermal isolation may be quite close to the adiabatic process, especially if it is a short process, since in this case no noticeable heat exchange between the gas and the surroundings takes place. This is the reason why a rapidly compressed gas is heated and a rapidly expanding gas cools down.

We can demonstrate this with the help of the following experiment. Take a glass bottle with a narrow neck and pour some water into it. Close the bottle with a rubber plug through which a tube connected to a compressor passes and begin pumping air into the bottle. Water will vanish in the process, pointing to the fact that the temperature inside the bottle rises. At a high enough pressure the plug is expelled from the bottle and fog appears inside it, this being an indication of a drop in temperature of the air in the process of its expansion.

Rapid forceful compression may result in a great rise in the temperature of a gas. If this gas consists of air and of vapours of petrol or of some other combustible substances, they will ignite. This phenomenon is utilized in the diesel engine for igniting the fuel-air mixture. (Why does a pump become hot when you pump tyres quickly?)

### 8-7 Some Ideas on Stellar Structure

The Sun is a typical star, that is, it is a gigantic ball of gas. The physical characteristics of the Sun are those of a medium star with a mass of  $2 \times 10^{30}$  kg and a radius of  $7 \times 10^8$  m.

The Sun consists mainly of hydrogen (about 70 per cent of its mass) and of helium (about 29 per cent). The mass of the Sun is 330 000 the Earth's mass and an enormous force of gravitation greatly compresses the gases. If one computes the volume of the solar sphere, he or she will find that the average density of the solar matter is  $1.4 \times 10^3$  kg/m<sup>3</sup>, that is, it exceeds the density of water. The pressure and density of the gas increase in the direction of the centre, reaching there the value of about  $1.5 \times 10^5$  kg/m<sup>3</sup> (over ten times the density of lead).

Every second the Sun radiates into space an enormous amount of energy. The source of the energy are the thermo-

nuclear reactions taking place inside the Sun. The temperature in the centre of the Sun is as high as  $13 \times 10^6$  K decreasing gradually with the distance from the centre.

Energy transport from inside the Sun is mainly effected by radiation. Upper layers absorb the radiation of internal layers and, in their turn, transmit energy to the layers more distant from the centre and so on until it reaches the layer whose radiation, at last, is able to penetrate into outer space.

The layer in which the visible radiation of the Sun is generated is termed *photosphere* and is observed as the solar disk. The photosphere is several hundred kilometres in thickness and the pressure in it is of the order of 0.1 atm. The temperature of the internal layers of the photosphere is about 6000 K decreasing to 4500 K in its outer layer. The photosphere is the lower layer of atmosphere of the Sun. The layer above it is called *chromosphere*, and the term for the outermost rarefied part of the atmosphere is called the *solar corona* (Fig. 8.3).

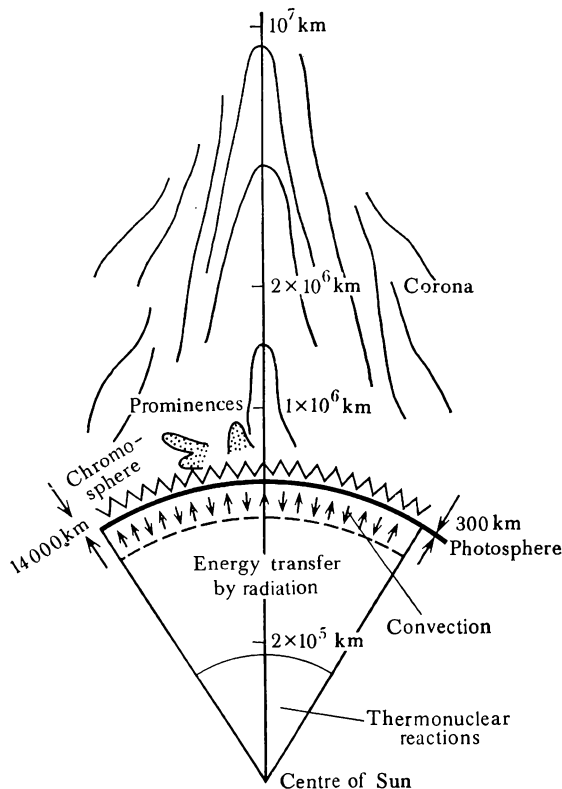
The gas of the chromosphere and corona is greatly rarefied: the upper layers of the chromosphere contain only about  $10^{15}$  atoms/m<sup>3</sup>, and the concentration in the corona is by an order of magnitude less (compare with the atmosphere of the Earth:  $N_L = 2.7 \times 10^{25}$  molecules/m<sup>3</sup>, see Section 3-6).

The gas in the photosphere loses its energy due to radiation into the outer space and rapidly cools down. Because of that vertical mixing, convection of gas, takes place in the underlying layer. When the photosphere is observed through a telescope, it is seen to consist of numerous *granulation cells*: small clouds of hot gas which rise from the depth forcing out cooled gas; after several minutes they disintegrate to be replaced by new ones. Sometimes rising stable fluxes of hot gases called *flares* are formed. They are visible as large patches of bright material. From time to time *sunspots*, regions of lower temperature, appear in the photosphere. By the way, the motion of sunspots helped to discover the rotation of the Sun.

Powerful convection fluxes of solar matter create mechanical vibrations and waves similar to sound waves. As those waves propagate to the upper layers of the atmosphere, where the gas is very rare, the amplitude of vibrations of gas particles increases to several kilometres and the velocities rise greatly. Such vibrations cannot, however, retain their regular character for a long time. Large masses of gas that take part in the wave motion disintegrate into individual small masses moving at random. The result is that the average energy of random motion of gas particles rises

greatly at the expense of the mechanical energy of the waves and the temperature in the chromosphere rises to several tens of thousands kelvins and in the corona to  $10^6$  K. However, owing to the very low density of the solar corona, its brightness is one million times less than that of the photosphere and does not exceed the brightness of the Moon. The corona can be observed during total eclipses of the Sun,

**Fig. 8.3** Schematic cut-away view of the Sun and its atmosphere.



when the Moon covers the bright disk of the photosphere. The structure of the corona is in the form of rays whose wavelength is sometimes ten times longer than the radius of the photosphere.

The kinetic energy of the gas particles in the solar corona is so great that many of them overcome the Sun's attraction and fly out into the interplanetary space. Fluxes of such particles flying at speeds of hundreds kilometres per second are called *solar wind*. Note that the pressure of the solar wind

and of light radiated by the Sun is responsible for the formation of comet tails, which are always turned away from the Sun.

## Change of State—I

## 9

### 9-1 Vapourization and Condensation

Let us turn now to the properties of a substance in various states of aggregation. The properties of gases are the easiest to study and they will be discussed first. It was mentioned above that many properties of the gases are independent of their nature. However, with a decrease in temperature and an increase in pressure the dependence of the properties on the nature of a gas becomes more noticeable. Under such conditions the gas is said to be a *vapour*, which underlines the fact that it originated from a liquid. At a further decrease in temperature and increase in pressure the gas becomes a liquid.

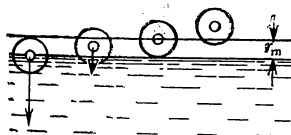
We shall see below that a more precise distinction may be drawn between a gas and a vapour. Here we would only like to note that the properties of a gas that is far away from the point of conversion into a liquid are similar to those of the ideal gas and are already known to us. Therefore below we shall discuss only the properties of vapours which manifest themselves in the processes of transition from the liquid state to the gaseous and back. We shall start our studies of the properties of vapours with such processes.

The transition of a substance from the liquid state to the gaseous state is termed *vapourization*, and the term for the transition from the gaseous state to the liquid state is *condensation*. We recall that the process of vapourization is intimately connected with the increase in internal energy of the substance, and the process of condensation with its decrease. Therefore vapourization and condensation can take place only in the course of heat exchange between the substance and the surrounding medium. In nature vapourization takes place in the forms of evaporation and boiling.

### 9-2 Evaporation

Vapourization which takes place only from the free surface of a liquid in contact with a gaseous medium or vacuum is termed *evaporation*. The drying of water in a saucer or of

**Fig. 9.1** Molecules leaving water have to overcome opposition of forces pulling them back into liquid.



a puddle on the floor are examples of evaporation of a liquid.

Let us discuss the process of evaporation from the point of view of the kinetic theory. The potential energy of the molecules of a liquid is known to increase with the increase in the intermolecular distances. Therefore, in order to leave the liquid, the molecule must perform work at the expense of its kinetic energy. There are always some molecules at the surface of the liquid that move at random and try to leave the liquid. When such a molecule leaves the surface, a force which pulls it back into the liquid (Fig. 9.1) appears. Therefore only molecules whose kinetic energy exceeds the work needed to overcome the molecular forces in a layer  $r_m$  thick leave the liquid (here  $r_m$  is the radius of molecular interaction).

The molecules which manage to leave the liquid make up the vapour above the surface of the liquid. Since only the molecules with sufficiently high kinetic energy are able to leave the liquid, the kinetic energy  $\bar{K}_{\text{trans}}$  of molecules remaining in it decreases in the process of evaporation, that is, the liquid cools down. This explains why we feel cold when we come out of water after bathing, the cooling of your hand if you wet it with ether, etc.

Some of the molecules of the vapour that move at random over the surface of the liquid return into it. This means that evaporation of a liquid is always accompanied by condensation of its vapour. The vapour molecules entering the liquid are incorporated into its structure. This means that  $\bar{K}_{\text{trans}}$  of the molecules of the liquid rises and hence the temperature of the liquid rises too. Thus there are two processes, evaporation and condensation, which take place on the surface of the liquid simultaneously. Prevailing evaporation causes the liquid to cool and prevailing condensation to heat.

Since the forces of molecular interaction depend on the nature of the molecules one may expect the rate of evaporation to depend on the type of the liquid. This can easily be established by experiments. Should equal volumes of different liquids be poured into identical open vessels, it will become evident after some time that the evaporation rate of those liquids is different. Ethyl ether evaporates more rapidly than alcohol, and alcohol more rapidly than water. Experience shows the evaporation rate of a liquid to depend on the area of its free surface. The greater the area the greater the evaporation rate of the liquid. (Cite examples to support this contention and explain it on the basis of the kinetic theory.)

It may easily be seen that the evaporation rate rises with the increase in temperature. For instance, hot water evapo-

rates at a greater rate than cold water. The explanation is that the rise in temperature brings about an increase in the average kinetic energy of the molecules in the liquid, and this leads to the rise in the number of molecules capable of surmounting the opposition of the liquid's surface layer and of flying beyond it.

As was noted above, the evaporation of a liquid is always accompanied by condensation of its vapour, leading to a decrease in the evaporation rate. Since the condensation rate increases with the increase in the density of vapour above the surface of the liquid, such an increase should result in a reduction in the evaporation rate. Experiments show that the evaporation rate of a liquid is indeed the greater the less the density of the vapour above its surface is. For instance, when we blow on hot water in a saucer, we reduce the density of the water vapour above it and thereby increase the evaporation of water. And this, in turn, makes it cool quicker. Finally, increasing the external pressure on the liquid's surface reduces its evaporation rate. (Explain why.)

Note in addition that it is not always possible to observe the cooling of a liquid in the process of its evaporation. The reason is that there is always some heat exchange between the liquid and the surrounding bodies which compensates its energy losses and thereby decreases its rate of cooling. However, despite this fact, if the evaporation rate of a liquid is high, its temperature can drop substantially. Therefore by using liquids with a high evaporation rate (alcohol, ethyl ether) one can obtain an appreciable reduction in temperature. For instance, this is used to reduce the sensitivity of the skin by wetting it with ether.

### 9-3 Heat of Vapourization

It was mentioned in Section 6-5 that vapourization at constant temperature results in an increase in the internal energy of the substance and condensation results in a decrease. Since the temperatures of liquid and vapour can be the same, all variations of the internal energy of the substance are due entirely to the variations of the potential energy of its molecules. Hence, if the temperatures are the same, a unit mass of a liquid has a smaller internal energy than a unit mass of its vapour.

Experiments show that in the process of vapourization the density of a substance decreases greatly and the volume occupied by the substance increases. Therefore vapourization



should be accompanied by work performed against external pressure. Therefore the energy which must be transferred to the liquid to turn it into vapour at a constant temperature is spent partially on increasing the internal energy of the substance and partially on performing work against external forces in the process of its expansion.

In practice, to turn a liquid into vapour, heat is supplied to it by heat exchange. The heat  $Q_v$  required to turn a mass of liquid into vapour at constant temperature is termed *heat of vapourization*. When a mass of vapour is to be turned into liquid, heat  $Q_v$  called *heat of condensation*, should be taken away from it. Under the same external conditions the heat of vapourization is equal to the heat of condensation.

It was established with the aid of a calorimeter that the heat of vapourization,  $Q_v$ , is directly proportional to the mass  $m$  of the liquid vapourized:

$$Q_v = rm \quad (9.1)$$

where  $r$  is a proportionality factor whose value depends on the nature of the liquid and on external conditions.

The quantity  $r$  characterizing the dependence of the heat of vapourization on the nature of the substance and on external conditions is termed *specific heat of vapourization*. Specific heat of vapourization is the heat required to convert a unit mass of liquid into vapour at constant temperature:

$$r = Q_v/m \quad (9.1a)$$

The unit for measuring  $r$  in the SI system is the specific heat of vapourization of such a liquid that requires 1 J to convert 1 kg of its mass into vapour. (Demonstrate it using formula (9.1a).)

As an example we note that the specific heat of vapourization of water at a temperature of 373 K (100°C) is  $2.26 \times 10^6$  J/kg.

Experiments show that the evaporation of a liquid takes place at all temperatures. For instance, water evaporates at 273 K (0°C) and at 373 K (100°C). Therefore there is always the vapour of a liquid in contact with its surface.

Since vapourization takes place at different temperatures, the question is: Does the specific heat of vapourization vary with temperature or does it remain the same at all temperatures? The fact is that it does vary. Experiments show that the specific heat of vapourization decreases with a rise in temperature. This is because all liquids expand upon heating with the result that the intermolecular distances increase and the molecular forces of interaction become smaller. Besides, the higher the temperature the greater  $\bar{K}_{\text{trans}}$  of

the liquid's molecules and the less additional energy is required to make them capable of flying beyond the surface.

Figure 9.2 shows the dependence of the specific heat of vapourization on temperature for two liquids (water and

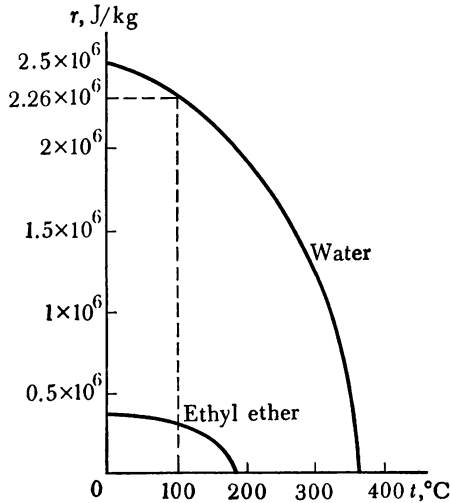


Fig. 9.2 Specific heat of vapourization for water and ethyl ether at different temperatures.

ethyl ether): Note that at first  $r$  decreases gradually with the increase in  $t$  but then drops rapidly to zero. There is a temperature at which  $r$  drops to zero for every liquid. Its physical meaning shall be made clear in Section 10-8.

## Properties of Vapour. Boiling

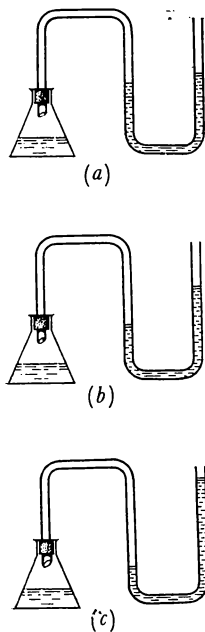
## 10

### 10-1 Nonsaturated and Saturated Vapours

When the free surface of a liquid in a vessel is in contact with an open atmosphere, evaporation exceeds condensation and the level of the liquid sinks in the course of time. This is because moving air carries vapour away with it, this leading to the decrease in the density of the vapour above the liquid's surface.

Experience shows that the level of a liquid in a hermetically sealed vessel remains constant. This means that in such a vessel the evaporation of the liquid is fully compensated by the condensation of its vapour, that is, the number of the molecules leaving the liquid is equal to the number return

**Fig. 10.1** Experiments demonstrating dependence of saturated vapour pressure on type of liquid: (a) water; (b) alcohol; (c) ethyl ether.



ing into it. In other words, the numbers of molecules both in the liquid and in the vapour above it remains unchanged despite the fact that the liquid and the vapour exchange molecules. Such an equilibrium between the liquid and its vapour is termed *dynamic equilibrium*.

A vapour in dynamic equilibrium with its liquid is termed *saturated vapour*. A vapour in contact with its liquid when evaporation exceeds condensation and a vapour in the absence of its liquid are termed *unsaturated*. It may easily be reasoned that for a definite temperature the maximum density and the maximum pressure will be those of a saturated vapour.

To see whether the density and pressure of a saturated vapour depends on the type of substance, let us make the following experiment. We take three closed bulbs containing water, alcohol and ethyl ether and connect them to pressure gauges (Fig. 10.1). The pressure in the bulbs will be due, besides air, also to saturated vapours of the liquids contained in them. The maximum pressure will be that in the bulb containing ether and the minimum in the bulb containing water, that is, the greatest pressure is set up by the saturated vapour of the liquid which evaporates at the highest rate. Experiments of that sort yielded the following result: the smaller the specific heat of vapourization of a liquid the greater is its evaporation rate and the greater are the pressure and the concentration of its saturated vapour (at equal temperatures).

## 10-2 Properties of Saturated Vapour

Let us see how saturated vapour behaves in the isochoric process. To this end we take a hermetically sealed vessel provided with a pressure gauge. Note that the liquid is poured into the vessel before it is hermetized and that the volume above the liquid contains only the vapour of this liquid. Next we place the vessel into a water bath (Fig. 10.2) and start heating it recording the temperature and the pressure of the saturated vapour in it. Then we stop heating and start cooling the vessel, again recording the temperature and the pressure inside. Comparing the readings of the pressure gauge at the same temperatures, we shall see that they are equal. This means that the pressure and the density of saturated vapour are uniquely determined by its temperature. The results of such experiments are presented in Table 10.1.

We see that the pressure of a saturated vapour depends on its nature and rapidly rises with temperature. Observing

**Table 10.1** Saturated vapour pressure for water, ethyl alcohol and ethyl ether (in mmHg) at different temperatures

$t, ^\circ\text{C}$	Water	Ethyl alcohol	Ethyl ether
0	4.6	13	186
10	9.2	24	290
20	17.5	45	440
30	31.8	79	640
35			760
40	55	134	920
50	92	220	1270
60	147	350	1740
70	232	560	
78		760	
80	353	830	
90	524	1200	
100	760	1670	
120	1520		

the level of the liquid inside the vessel in the course of the experiment, we shall notice that it sinks upon heating and rises upon cooling. This means that the mass and the density of the vapour inside the vessel rise upon heating and fall upon cooling. The aforesaid leads to the conclusion that there are two causes for the increase in the pressure of saturated vapour upon heating: firstly, the increase in  $\bar{K}_{\text{trans}}$  of the vapour molecules and, secondly, the increase in their number per unit volume of vapour, that is, the increase in its density.

Note that when the ideal gas is heated, only the first cause is responsible for the increase in pressure since the mass of the gas and its density remain constant in this process. Figure 10.3 shows a typical plot of the temperature dependence of saturated vapour pressure (curve *a*), and the plot below is the same dependence for the ideal gas in an isochoric process (straight line *b*) shown for the sake of comparison.

It follows from the experiments discussed above that in an isochoric process saturated vapours do behave as the ideal gas. The principal explanation is that the mass of saturated vapour changes during an isochoric process.

Consider now the isothermal process. To this end take a cylindrical vessel filled with a small amount of liquid, the design of the vessel being the same as is the previous experiment the only difference being that it contains a mobile piston (Fig. 10.4*a*). Should we move the piston downwards

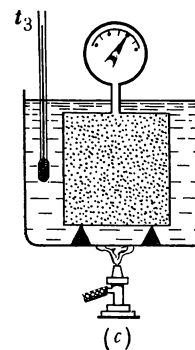
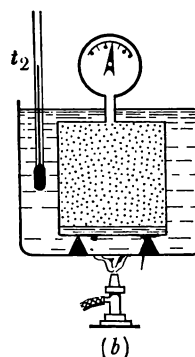
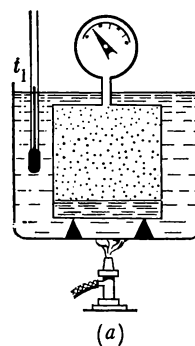
**Fig. 10.2** Experimental arrangement for demonstrating dependence of vapour pressure on temperature at constant volume.

Fig. 10.3 Temperature dependence of vapour pressure (curve *a*) and respective plot for ideal gas (curve *b*); upper section of curve *a* corresponds to nonsaturated vapour.

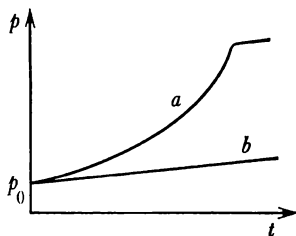
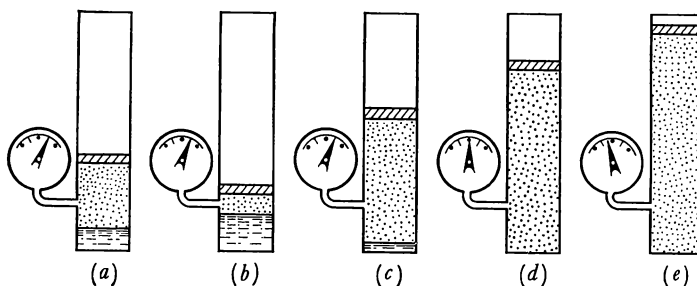


Fig. 10.4 Experimental arrangement demonstrating dependence of vapour pressure on volume at constant temperature. The saturated vapour pressure is independent of volume (*a*, *b*, *c*); when liquid evaporates completely the vapour becomes nonsaturated and its pressure begins to change with volume (*d*, *e*).



volume of saturated vapour. Hence the density of saturated vapour remains unchanged in an isothermal process. This supports the statement made above that the pressure and density of a saturated vapour depend only on the temperature and the nature of the substance.

It may be concluded from the aforesaid that the laws of the ideal gas are not valid for saturated vapours. The reason is that in any process involving a saturated vapour its mass is subject to change.

### 10-3 Properties of Nonsaturated Vapour

Should the vessel containing a liquid shown in Fig. 10.2 be heated after the liquid in it has disappeared (Fig. 10.2*c*) the vapour would become unsaturated. Its density would remain constant with further heating (explain why) and the rise in its pressure with the rise in temperature would not be so rapid (Fig. 10.3, curve *a*). However, not far away from saturation the effect of molecular interaction would still be noticeable and only after substantial heating would the

unsaturated vapour behave like the ideal gas in an isochoric process.

In the course of isothermal expansion described in the preceding section we shall be able to note the change in the pressure of the vapour when it becomes unsaturated (Fig. 10.4*d* and *e*). As long as the density of unsaturated vapour is close to that of saturated vapour the effect of molecular interaction and of their proper volume remains great and the dependence of the pressure of the vapour on its volume does not obey the Boyle law. Hence one may apply the ideal-gas law to unsaturated vapours only if they are far from saturation.

Analyzing the conclusions of two preceding sections we easily establish that a saturated vapour can be converted into an unsaturated either by isochoric heating or isothermal expansion, or by simultaneously both. In the same manner, an unsaturated vapour can always be converted into a saturated either by isochoric cooling or isothermic compression, or by simultaneously both.

The proper volume of the molecules of a vapour is practically always negligible in comparison with the volume occupied by the vapour. Therefore the presence in a volume of the vapour of some other liquid (even if the vapour is saturated) does not obstruct the evaporation of the first liquid. In this case the combined pressure of the vapours will be equal to the sum of the individual (partial) pressures. This property of the vapours is expressed by means of the law discovered by the English chemist John Dalton (1766-1844) in the absence of chemical interaction between vapours or gases the pressure of a mixture of them is equal to the sum of the partial pressures of each.

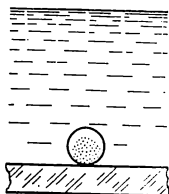
#### 10-4 The Boiling Process

Another form of vapourization is the boiling of liquids. It was established by experiments that while a liquid boils its temperature remains constant. Vapourization which takes place throughout the volume of a liquid at a constant temperature is termed *boiling*.

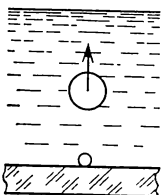
Let us study the particulars of the boiling process in a liquid. Pour water into a glass bulb and observe it in the process of heating. As the temperature of the water rises, gas bubbles appear on the bottom and walls of the bulb. Let us see how this happens.

The surface of a solid is capable of retaining gas molecules which, so to say, stick to it. Such "sticking" of gas mole-

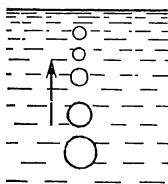
Fig. 10.5 Boiling of liquid: (a) bubble at bottom of vessel; (b) bubble detaches itself and leaves nucleus for new bubble; (c) volume of bubble decreases as bubble rises; (d) in boiling, volume of bubble increases as it rises.



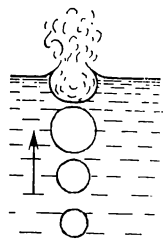
(a)



(b)



(c)



(d)

cules to the molecules of the solid's surface, is termed *adsorption* and the gas bonded with the solid's surface is termed *adsorbed*.

Experiments show that gases are soluble in liquids and that the solubility of a gas in a liquid decreases with the rise in temperature. Because of that, when water is heated the air dissolved in it precipitates on the walls and joins the air adsorbed on them.

When water is heated the number of bubbles grows and so does their volume. Since the bubbles are in water, they contain water vapour in addition to air. The growth of the bubbles in the process of heating is due to the increase in the volume of air they contain as well as to the increasing pressure of the water vapour. The following factors obstruct the expansion of the bubbles: external pressure of the atmosphere on the liquid's surface, hydrostatic pressure of the column of water of the height equal to the distance of the bubbles from the liquid's surface, and the pressure due to the curvature of the bubble's surface and proportional to the bubble's radius (see Section 12-6). For very small bubbles the Laplace pressure may exceed the atmospheric, but for large ones it may be neglected. The pressure due to curvature drops as the bubble grows and this facilitates the swelling of the bubbles.

When the volume of the bubble is large enough, the Archimedes force tears it away from the bottom or from the wall, and it rises to the surface leaving a nucleus for a new bubble (Fig. 10.5a and b). When a liquid is heated from below its upper layers are colder than the lower, the water vapour in the rising bubble condenses and the air contained in it begins to dissolve in water again. Thus the volume of the bubble begins to decrease. Many bubbles fail to reach the surface of the water and vanish. Some reach the surface, but the amounts of air and of water vapour remaining in them are by that time quite small. The remaining air and water vapour go out into the surrounding medium (Fig. 10.5c). This continues until convection equalizes the temperature throughout the liquid.

With the temperature being equal throughout the liquid the rising bubbles will grow in volume. The explanation is as follows. As a bubble rises in a liquid in which the temperature is everywhere the same the saturated vapour pressure inside it remains constant, but since the hydrostatic pressure drops it grows in dimensions. Since the pressure of saturated vapour (steam) is independent of the volume as the bubble grows the entire volume inside it is filled with saturated steam (Fig. 10.5d). Note that in the course of the bubble's

growth the pressure due to curvature decreases somewhat and this, too, facilitates the growth of the rising bubble.

When such a bubble reaches the surface of the liquid, the pressure of saturated steam inside it is practically the atmospheric pressure at the surface since the hydrostatic pressure is then zero and both the pressure of air inside and the pressure due to curvature are negligible. On the liquid's surface the bubble bursts and the substantial amount of saturated steam inside it is released into the atmosphere. The process of growth of bubbles containing saturated steam and of release of this steam into the atmosphere is in fact what we call boiling. Thus, boiling of a liquid takes place at an equal temperature throughout the liquid when the pressure of the saturated vapour of this liquid is equal, or exceeds, the external pressure.

Experiment proves the temperatures of the boiling liquid and of the vapour above its surface to be equal. This means that the entire energy supplied to a liquid in the process of its boiling is spent on increasing the potential energy of its molecules and on the work against external forces performed by the expanding substance.

In view of the aforesaid the following definition could be given: the *boiling temperature* of a liquid is the temperature at which the pressure of saturated vapour of this liquid is equal to the external pressure at its surface.

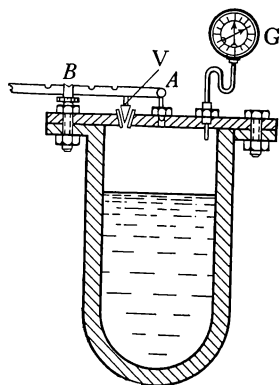
### 10-5 Dependence of Boiling Temperature on External Pressure. Boiling Point

Since the pressure of the saturated vapour is uniquely determined by the temperature and since boiling of a liquid starts at the moment the pressure of the saturated vapour of this liquid becomes equal to the external pressure, the boiling temperature must be dependent on the external pressure. It may easily be demonstrated with the aid of experiments that a reduction in the external pressure decreases the boiling temperature and an increase in the external pressure leads to its rise.

The boiling of a liquid at a pressure below atmospheric may be demonstrated with the aid of the following experiment. A glass is filled with water from the water pipe and a thermometer is immersed in it. The glass with water is placed under a bell and the pump is turned on. When the pressure becomes low enough, the water in the glass begins to boil. Since vapourization requires energy, the temperature



Fig. 10.6 Schematic representation of autoclave: V, safety valve; AB, lever pressing valve; G, pressure gauge.



in the glass will fall in the process of boiling and, if the pump is effective, the water will finally freeze.

Heating water to high temperatures is effected in *boilers* and *autoclaves* (Fig. 10.6). At pressures above 100 atm water may be heated to temperatures in excess of 300° C.

The boiling temperature of a liquid at standard atmospheric pressure is termed the *boiling point* (see Table 10.2). It follows from Tables 10.1 and 10.2 that the saturated vapour pressures of water, ethyl alcohol and ethyl ether at their boiling points is  $1.013 \times 10^5$  Pa, or 760 mmHg.

The obvious conclusion is that in deep mines water should boil at temperatures above 100 °C and in mountainous areas at temperatures below 100 °C. Since the boiling temperature of water depends on the altitude above sea level, the thermometer scale may be graduated at the altitude at which water boils at some temperature instead of the temperature itself. The term for the method of measuring altitude with the aid of such a thermometer is *hypsometry*.

Experiments show that the boiling temperature of a solution is always higher than that of the pure solvent and that it rises with the solution's concentration. However, the temperature of the vapour above the boiling solution is equal to the boiling temperature of the pure solvent. Therefore, for measuring the boiling temperature of a pure liquid, it is advisable not to immerse the thermometer into the liquid but to place it above the surface of the boiling liquid in its vapour.

The boiling process is closely related to the presence of gas dissolved in the liquid. If this gas is withdrawn by means of, say, prolonged boiling, it will be possible to heat such a liquid to temperatures higher than its boiling point. The term for such a liquid is *superheated*.

In the absence of gas bubbles the nucleation of tiniest bubbles of vapour which could become centres of vapourization is hampered by the pressure due to curvature which is quite large for bubbles of small radius. This explains the phenomenon of a superheated liquid. When such a liquid finally starts boiling it does so very intensely.

Table 10.2 Boiling points of substances

Substance	$t_b$ , °C
Acetone	56.2
Benzene	80
Ethyl alcohol	78
Ethyl ether	35
Mercury	357
Water	100

### 10-6 The Law of Heat Exchange for Vapourization and Condensation

Let us now turn to calculating the heat required to transform a liquid into vapour by boiling it. Since the liquid boils at the boiling point,  $T_b$ , it should be first heated from the initial temperature  $T_1$  to the boiling point  $T_b$  and then

evaporated. Figure 10.7 shows the dependence of  $T$  on the heat  $Q$  supplied to the liquid.

The heat  $Q_{\text{liq}}$  required to bring the liquid to its boiling point is found from the formula

$$Q_{\text{liq}} = c_{\text{liq}} m (T_b - T_1)$$

where  $m$  is the mass of the liquid, and  $c_{\text{liq}}$  is its specific heat. To calculate the heat  $Q_v$  required to convert the liquid into vapour without changing the temperature the following formula is used:

$$Q_v = rm$$

Hence, the total heat is expressed by the relation

$$Q = c_{\text{liq}} m (T_b - T_1) + rm$$

The specific heat of vapourization of a substance is found from experiment with the aid of the law of heat exchange. Let us show how this is done using water as an example. To this end we take a calorimeter filled with water at a temperature  $T_1$ . Vapour (steam) at a temperature of 373 K is delivered from a boiler through a tube immersed in cold water contained in the calorimeter and condenses there. After some time the tube is withdrawn and the final temperature  $\Theta$  is measured. Next the mass of condensed vapour is determined by weighing and the law of heat exchange is written.

In this experiment the recipients of heat are the calorimeter and the cold water contained in it:

$$\begin{aligned} Q_{\text{gained}} &= Q_{\text{cal}} + Q_{\text{water}} \\ &= c_{\text{cal}} m_{\text{cal}} (\Theta - T_1) + c_{\text{water}} m_{\text{water}} (\Theta - T_1) \end{aligned}$$

The losers of heat are the condensing steam and water obtained from it when cooled from  $T_b$  to  $\Theta$

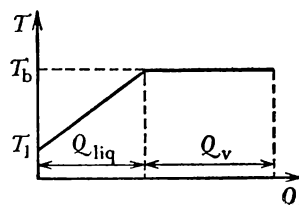
$$Q_{\text{lost}} = rm_v + c_{\text{water}} m_v (T_b - \Theta)$$

Since  $Q_{\text{lost}} = Q_{\text{gained}}$ , we have

$$rm_v + c_{\text{water}} m_v (T_b - \Theta) = (c_{\text{cal}} m_{\text{cal}} + c_{\text{water}} m_{\text{water}}) (\Theta - T_1)$$

This equation can now be solved for  $r$ .

Fig. 10.7 Dependence of temperature of liquid on heat supplied to it.



### 10-7 Superheated Steam and Its Use in Technology

Since a lot of energy is spent in converting water into steam, much work can be performed and much heat liberated in the process of its cooling and condensation. This energy is wide-

ly used to drive steam turbines operating in power stations, ships, and other devices.

Steam turbines with a power output over one million kilowatts are being built in the Soviet Union. Steam energy is being also used in industry (steam hammers, for instance), for heating homes, and other purposes.

Steam produced in a boiler is heated to high temperatures with the aid of special devices and then supplied to steam turbines. The term for such steam is *dry*, or *superheated*, steam. Since the pressure of steam rises simultaneously with its temperature, another term for superheated steam is *high-pressure steam*. The temperature of high-pressure steam is so high that the pipeline and turbine blades operating on such steam glow red. The pressure of steam can be as high as 300 atm, the turbine efficiency in this case being 40-45 per cent (the efficiency is the higher the higher the steam temperature).

After performing work in the turbine the steam still has a high temperature and substantial energy. At heat and electric power plants spent steam is supplied through special pipelines to factories and homes where the residual energy of the steam is utilized. Such a system makes possible a better utilization of energy of the fuel at such plants. The Soviet Union is foremost among the countries of the world in the development of centralized heating.

### 10-8 Critical State of Substance

It was explained in Section 9-1 that to convert vapour into liquid one should raise the pressure and lower the temperature. Using this method the English chemist and physicist Michael Faraday (1791-1867) succeeded in converting into liquid state many substances which before were known to exist only in the gaseous state. However, for a long time the efforts to liquefy some of the gases proved unsuccessful despite great pressures used. The Russian chemist Dmitri Mendeleev (1834-1907) gave the theoretical explanation for this.

The boundary separating the liquid from the surrounding medium is the free surface of the liquid. The existence of such a boundary enables us to state definitely where the liquid phase is and where the gaseous phase is. Such a pronounced difference between the liquid and its vapour is mainly due to the fact that the density of a liquid is many times that of its vapour. However, when a liquid is heated in a sealed vessel, its density decreases and that of the vapour above the surface increases. This means that the

difference between the liquid and its saturated vapour becomes less pronounced and should vanish altogether at a high enough temperature.

In 1861 Mendeleev proved that every liquid must have a temperature at which there is no difference between the liquid and its vapour. Mendeleev's term for it was "temperature of absolute boiling". But it was the Irish chemist

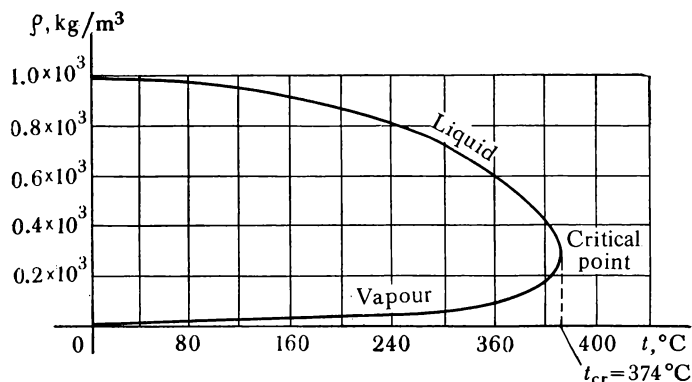


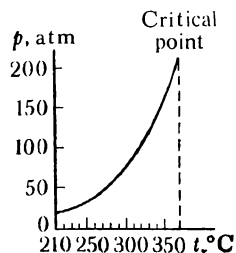
Fig. 10.8 Temperature dependence of densities of water and its saturated vapour.

Thomas Andrews (1813-1885) who studied the process of transformation of vapour into liquid and back at various pressures experimentally. He proved the existence of such a temperature for every liquid and introduced a new term for it, *critical temperature*, which is used at present.

The critical temperature is the temperature at which the density of a liquid becomes equal to that of its saturated vapour. The plot of the temperature dependence of the densities of water and its saturated vapour is shown in Fig. 10.8; it follows that the critical temperature,  $t_{cr}$ , for water is 374 °C. Since not only the density of saturated vapour but its pressure as well are uniquely determined by its temperature, one may plot the  $p$  versus  $t$  dependence for saturated vapour (Fig. 10.9).

The saturated vapour pressure of a substance at its critical temperature is termed critical pressure,  $p_{cr}$ . It is the maximum pressure of the saturated vapour of a substance. Its value for water is 218.5 atm. It may be seen from Fig. 9.2 that at the critical temperature the specific heat of vapourization of water is zero. This is also true for other liquids. Therefore at the critical temperature all differences between the vapour and the liquid vanish together with the boundary separating them. This means that at temperatures above  $t_{cr}$  a substance exists only in one, gaseous, state and no increase in pressure will turn it into a liquid.

Fig. 10.9 Pressure of saturated water vapour versus temperature (the plot terminates at critical point).



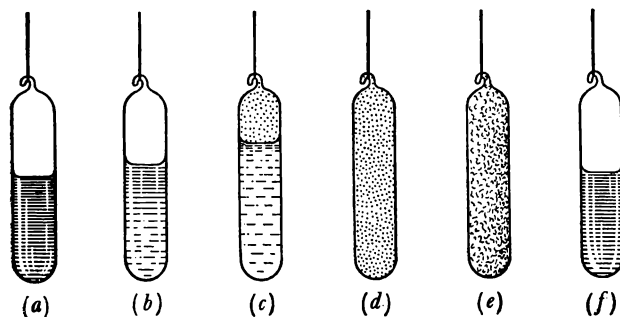
A substance at critical temperature and critical pressure is said to be in the *critical state*. The volume occupied by a substance in the critical state is termed the *critical volume*,  $V_{cr}$ . It is the maximum volume the available mass of the substance can occupy remaining in the liquid state. Tables usually contain values of critical volume per mole of the substance. The values of  $t_{cr}$ ,  $p_{cr}$  and  $V_{cr}$  for a substance are termed *critical constants of the substance* (see Table 10.3).

Table 10.3 Critical constants for some pure substances.

Substance	$t_{cr}$ , °C	$p_{cr}$ , atm
Water	374.2	218.5
Ethyl alcohol	243.1	63.0
Ethyl ether	193.8	35.6
Oxygen	-118.4	49.7
Argon	-122.4	48.0
Nitrogen	-147.1	33.5
Neon	-228.7	26.9
Hydrogen	-241	12.8
Helium	-267.9	2.25

One may observe a substance passing through the critical state if one heats an ampoule of ethyl ether (Fig. 10.10).

Fig. 10.10 Heating ethyl ether in an ampoule to critical temperature ((a), (b), (c), (d)) makes it possible to observe transition through critical state (d); when temperature falls below critical, ether turns into liquid ((e), (f)).



Before sealing the ampoule is filled with a mass of ether whose volume in the critical state is equal to the internal volume of the ampoule.

We see that there is no principal difference between gas and vapour. The term *gas* usually applies to a substance in gaseous state when its temperature is above the critical. The term *vapour* also applies to the substance in the gaseous

state but when its temperature is below the critical. Therefore, vapour can be converted into liquid solely by increasing pressure. This cannot be done with a gas.

### 10-9 Liquefaction of Gases

When it was established that any gas can be liquefied if its temperature is below the critical, people using lower and lower temperatures gradually obtained all gases in liquid state. The last gas to be liquefied was helium, in 1908.

In gas liquefaction plants the gas is cooled in the process of its adiabatic expansion (see Section 8-6). Initially the gas is compressed to a high pressure in a compressor. The heat liberated in the process is transported away by a coolant (water). When subsequently the gas itself performs work in the process of adiabatic expansion (at the expense of its internal energy) its temperature drops substantially. The part of the plant where the gas expands performing external work (for instance, by displacing a piston) is termed *gas-expansion machine*.

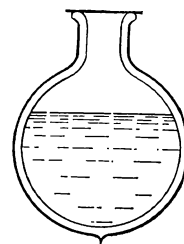
A notable contribution to the development of methods of gas liquefaction was made by the Soviet physicist P. L. Kapitza. In one of his plants using the turbine gas-expansion machine a jet of compressed gas is directed against the blades of a turbine; the gas drives the turbine and thus performs work and cools down.

Note that a highly compressed real gas, as distinct from the ideal gas, cools down upon expansion, even if it performs no work but simply flows out of a narrow nozzle. The explanation is that the expansion involves an increase in the intermolecular distances and this requires "internal" work against forces of molecular attraction to be performed by the molecules at the expense of their kinetic energy. This brings down their temperature. This method is also used for gas liquefaction.

When the temperature of the gas drops below the critical, it turns into a liquid. The liquefied gas is collected and stored in special Dewar vessels. For better heat insulation the vessels are made double walled with high vacuum between the walls to reduce heat conductivity (Fig. 10.11); to reduce heating of the liquid by radiation a film of amalgam is deposited on the walls (same as in mirrors).

The boiling points of various liquids in open Dewar vessels are given in Table 10.4. (Why is it not advisable to seal vessels containing liquefied gases?)

Fig. 10.11 Dewar vessel.



**Table 10.4** Boiling points of some liquefied gases

Substance	$t_b$ , °C
Xenon	—108
Krypton	—153
Oxygen	—183
Argon	—186
Nitrogen	—196
Neon	—246
Hydrogen	—253
Helium	—269

Air liquefaction is widely used for separation of the constituent gases. When liquid air boils, the more volatile gases (with lower boiling points, see Table 10.4) evaporate first. Nitrogen evaporates before oxygen and because of that after some time practically only pure oxygen remains in the vessel. It is used in metallurgy, as explosives, as rocket fuel, etc.

Air also contains some argon, helium and other inert gases. Since their boiling points are different, they can be separated with the aid of a special plant, a fractionating column, from liquid air cooled down to temperatures below the boiling points of the gases to be produced.

Liquefied gases are widely used in industry and in research for deep cooling of various substances. Many properties of substances experience notable changes at low temperatures, for instance lead becomes elastic and rubber becomes brittle. To obtain very low temperatures liquid hydrogen or helium boiling at a pressure below the atmospheric are being used. In the latter case a temperature of about one kelvin can be sustained. Research into the properties of substances at ultralow temperatures led to the discovery of superconductivity (see Section 18-10).

## II Water Vapour in the Atmosphere

### 11-1 Humidity

Since a continuous evaporation of water takes place from the surface of oceans, seas, lakes and rivers, the Earth's atmosphere always contains water vapour. It was established that the annual mass of water evaporating into the Earth's atmosphere is about  $4.25 \times 10^{14}$  tonnes, about one-fourth of this water falling out in the form of rain, snow or hail on land. Naturally, the amount of water vapour contained in the air is not everywhere the same. The air in the proximity of seas and oceans is more humid than deep inside the continents.

The quantity characterizing the contents of water vapour in air in different parts of the Earth's atmosphere is termed *humidity* of the air.

Humidity is of great importance for many processes taking place on Earth, for instance, for the evolution of the flora

and the fauna. Humidity greatly affects the crops and the productivity of animal breeding. No less important is humidity for many fields of technology, for instance, for drying processes, for storage of finished goods, etc. Therefore it is very important to be able to measure and control humidity.

## 11-2 Absolute and Relative Humidities

The parameters used for qualitative assessment of humidity are absolute and relative humidities.

The measure for *absolute humidity* of air is the density of water vapour  $\rho_{\text{abs}}$  contained in the atmosphere, or its pressure  $p_{\text{abs}}$ .

The relative humidity,  $B$ , presents a clearer picture of humidity.

The measure for *relative humidity* is the ratio of the absolute humidity,  $\rho_{\text{abs}}$  to the density of water vapour corresponding to saturation at the given temperature,

$$B = \frac{\rho_{\text{abs}}}{\rho_{\text{sat}}} \times 100\% \quad (11.1)$$

Relative humidity can also be found from the vapour pressure, for it is practically proportional to the vapour's density. Accordingly, one may also define  $B$  as follows: the measure for relative humidity is the ratio of the absolute humidity,  $p_{\text{abs}}$ , to the pressure of water vapour corresponding to saturation at the given temperature,  $p_{\text{sat}}$ :

$$B = \frac{p_{\text{abs}}}{p_{\text{sat}}} \times 100\% \quad (11.1a)$$

Hence, the relative humidity depends not only on absolute humidity but on the temperature as well. When calculating relative humidity one should find the values of  $\rho_{\text{sat}}$  or  $p_{\text{sat}}$  from tables. The density and pressure of saturated water vapour in the temperature range from 0 to 30 °C are shown in Table 11.1.

Let us discuss the effect of temperature variations on humidity. We assume that absolute humidity  $\rho_{\text{abs}}$  is 0.0094 kg/m<sup>3</sup> at 21 °C. Since the density of saturated water vapour  $\rho_{\text{sat}}$  at 21 °C is 0.0183 kg/m<sup>3</sup> (see Table 11.1), the relative humidity  $B$  will be about 50 per cent.

Assume now that the temperature drops to 10 °C and  $\rho_{\text{abs}}$  remains unchanged. Then the relative humidity will be 100 per cent, that is, the air will be saturated with water vapour.

**Table 11.1** Pressure and density of saturated water vapour at different temperatures

$t, ^\circ\text{C}$	$p_{\text{sat}}, \text{mmHg}$	$\rho_{\text{sat}}, \text{kg/m}^3$
0	4.6	0.0048
1	4.9	0.0052
2	5.3	0.0056
3	5.7	0.0060
4	6.1	0.0064
5	6.6	0.0068
6	7.0	0.0073
7	7.5	0.0078
8	8.0	0.0083
9	8.6	0.0088
10	9.2	0.0094
11	9.8	0.0100
12	10.5	0.0107
13	11.2	0.0114
14	12.0	0.0121
15	12.8	0.0128
16	13.6	0.0136
17	14.5	0.0145
18	15.5	0.0154
19	16.5	0.0163
20	17.5	0.0173
21	18.7	0.0183
22	19.8	0.0194
23	21.1	0.0206
24	22.4	0.0218
25	23.8	0.0230
26	25.2	0.0244
27	26.7	0.0258
28	28.4	0.0272
29	30.0	0.0287
30	31.8	0.0303



But if the temperature drops to  $7^{\circ}\text{C}$  (say, at night), 0.0016 kg of water vapour would condense from each cubic meter of air and dew will appear on grass. The temperature at which air containing a specified amount of water vapour becomes, upon cooling, saturated with water vapour is termed *dew point*. In the example cited the dew point is  $10^{\circ}\text{C}$ .

Note that, if dew point is known, the absolute air humidity may be found from Table 11.1, since  $\rho_{\text{abs}}$  is equal to the density of the saturated vapour  $\rho_{\text{sat}}$  at this temperature.

Clean air may be cooled down to a temperature below dew point without condensation of the water vapour in it. Such a vapour is termed *supersaturated*. The explanation for this phenomenon is that to convert vapour into liquid *condensation centres* are required, which are usually provided by dust particles. It was established that water vapour condenses on them more readily if the particles are charged. Individual ions can also serve as condensation centres.

### 11-3 Measuring Humidity

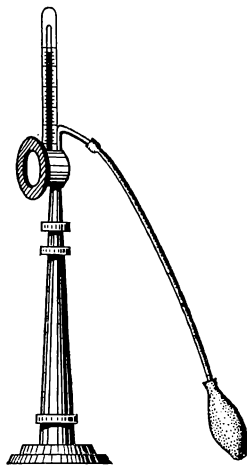
The practical instrument used to measure absolute humidity is the condensation hygrometer. Its principle of operation is based on the determination of the dew point, which is used to find absolute humidity from the tables (see Section 11-2).

The *condensation hygrometer* (from the Greek *hygros* for moist) consists of a circular metal box with a polished flat surface mounted on a support (Fig. 11.1). The box has two holes drilled in its upper part. One is used to fill the box with ether and to insert a thermometer and the other is connected to a pump or to a rubber bulb.

When air is blown through the ether, the box cools rapidly and dew appears on its surface, the polished surface turning mat. The dew point is determined by the thermometer's reading. For precise determination of the moment dew appears the polished surface is surrounded by a bright metal ring thermally insulated from the box. The mat surface of the box is clearly visible against the bright ring.

The *hair hygrometer* consists of a grease-freed hair, whose one end is fastened to a support and the other end is slung over a small pulley (Fig. 11.2). To maintain constant tension in the hair a small weight is attached to its free end. The operation is based on the fact that the hair elongates in moist air and contracts in dry air. The same property is peculiar also to a thin capron thread often used instead of a hair. When air humidity changes, the hygrometer's point-

Fig. 11.1 Condensation hygrometer.



er moves across its scale graduated with the aid of a standard instrument.

The *psychrometer* (from the Greek *psychros* meaning cold) consists of two identical thermometers. One thermometer is termed *dry* and the other *wet* (Fig. 11.3). The point of the wet thermometer is enclosed in muslin and is immersed in a water bath. Water evaporating from the muslin cools the thermometer's point. Because of that the readings of the wet thermometer are lower than those of the dry thermometer. The difference between those readings is the greater the drier the air. The relative humidity is obtained from the readings of the thermometers with the help of special tables supplied with the psychrometer. (In what case will the readings of both thermometers be the same?)

#### 11-4 The Atmosphere of Planets

The Earth is surrounded by an atmosphere—a layer of air held close to the surface of the Earth by gravitational forces. The atmosphere of the Earth is a mixture of nitrogen (78 per cent), oxygen (21 per cent) and small amounts of carbon dioxide, water vapour and of other gases.

The atmosphere plays a very important part in the heat balance of the Earth. It easily transmits the visible light of the Sun. The surface of the Earth absorbs this light, is heated by it and emits thermal radiation. However, water vapour and carbon dioxide contained in the atmosphere intensively absorb thermal radiation and thereby prevent the cooling of the Earth. In this way the atmosphere creates the so-called *greenhouse effect* and levels out 24-hour and seasonal temperature variations.

The pressure and density of the atmosphere fall off rapidly with altitude: the atmospheric pressure at an altitude eight kilometres is about three times smaller than at the sea level. The composition of the atmosphere at great altitudes changes, too, the contents of lighter gases such as helium and hydrogen rising. (Why?)

The atmospheres of other planets are quite unlike the Earth's. More extensively studied are the atmospheres of the planets of the Earth group, Venus and Mars. Already Lomonosov observing Venus with the Sun as background established that it is surrounded by a powerful atmosphere. Research by means of automatic space probes produced the result that the atmosphere of Venus consists of carbon dioxide (~ 97 per cent) and small amounts of water vapour, nitrogen and other gases. The surface of Venus is always covered

Fig. 11.2 Model of hair hygrometer.

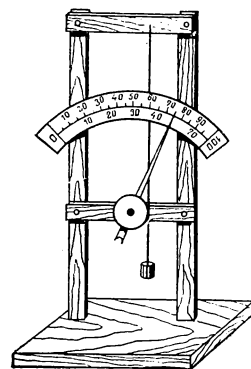
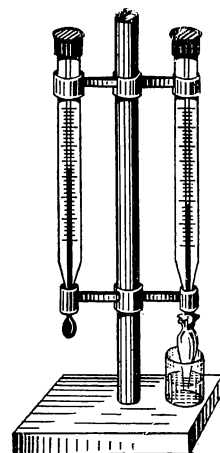


Fig. 11.3 Psychrometer



by a thick blanket of clouds. The atmospheric pressure at the surface is about 100 atm. Obviously, such a composition of the atmosphere and its high density favours the greenhouse effect. Therefore the temperature at the surface of Venus is very high—about 500 °C.

The atmosphere of Mars (whose mass is approximately one-ninth of the Earth's) consists also mainly of carbon dioxide gas (95 per cent), but its rarefaction is 200 times that of the Earth's and it is very dry. Therefore 24-hour temperature variations at Martian surface exceed 100 °C (the daytime temperatures on the equator being +30 °C and the nighttime temperatures -100 °C). The average temperature on the Mars is substantially below that on the Earth because Mars is farther away from the Sun than our planet.

The nature of the gigantic planets is quite different from that of the planets of the Earth group. Their masses are several tens of times greater than that of the Earth\* and they are surrounded by very powerful and thick atmospheres consisting of hydrogen, its compounds (ammonia, methane, etc.) and helium.

## 12

# The Liquid State

### 12-1 What Is a Liquid?

A substance in the liquid state is known to retain its volume and to assume the shape of the vessel which contains it. Let us see how these facts can be explained from the viewpoint of the kinetic theory.

The fact that a liquid retains its volume proves that there are forces of attraction between its molecules. Therefore the intermolecular distances in a liquid should be shorter than the radius of molecular interaction. Hence, if one draws a sphere of molecular interaction around a molecule, there will be several centres of other molecules inside this sphere and these will be the molecules that interact with our molecule. Those forces of interaction retain a molecule of the liquid in the vicinity of its temporary equilibrium position for about  $10^{-12}$ - $10^{-10}$  seconds after which it jumps to a new temporary equilibrium position which is at a distance of about its diameter from the initial position. Between the

\* The principal physical characteristics of the planets are presented in Table 43.1.

jumps the molecules of the liquid take part in vibrational motion about their temporary equilibrium positions. The time interval between two successive jumps of the molecule from one position to another is termed *time of residence*. This time depends on the nature of the liquid and on temperature. When the liquid is heated, the time of residence decreases.

During the time of residence (of the order of  $10^{-11}$  s) most of the molecules of the liquid are retained in their equilibrium positions and only some of them manage to change their positions. During greater time intervals most of the molecules will change their positions and this is the reason for the fluidity of a liquid and for its tendency to assume the shape of the vessel that contains it.

Since the molecules of a liquid are rather close to each other, a molecule which receives a high kinetic energy, even if it is able to overcome the attraction of its nearest neighbours, will find itself inside the sphere of interaction of other molecules and will occupy a new equilibrium position. Only molecules close to the surface of the liquid are able to leave the liquid, this being the explanation for the evaporation process.

Should we observe a very small element of volume of a liquid for a time equal to the time of residence, we would notice that the arrangement of molecules in it is regular just as in the crystal lattice of a solid. After some time this arrangement breaks up to be re-established at some new place. In this sense all the volume occupied by the liquid may be regarded as made up of numerous nuclei of crystallization which, however, are unstable, that is break up in some places and are born anew in others.

Hence the structure in a small volume of a liquid is regular, but it is chaotic in a large volume. In this sense a liquid is said to possess a short-order molecular structure, but to have no long-order structure. Such a structure of a liquid is termed quasi-crystalline (crystal-like). Note that at high enough temperatures the time of residence becomes quite small and the short-order structure in the liquid practically disappears.

A liquid can display properties peculiar to a solid. If a force acts on the liquid in a small enough time interval, the liquid displays elastic properties. For instance, if one strikes the surface of water swiftly with a stick, the stick may be forced out of his hand or even break; a stone can be thrown so that it recoils from the surface of water and only after several such leaps sinks into it. If, on the other hand, the time a liquid is acted upon is great, instead of elasticity

the property of *fluidity* is displayed. For instance, a man can easily submerge his hand in water.

When a force acts during a short time on a jet of a liquid, the latter exhibits *brittleness*. The tensile strength of a liquid although not so great as that of a solid is not much smaller. For water it is  $2.5 \times 10^7 \text{ N/m}^2$ . The *compressibility* of liquids is also quite small, although it exceeds the values for appropriate solids. For instance, an increase in pressure of 1 atm reduces the volume of water by 50 parts in a million.

Ruptures inside a liquid free from foreign objects, for instance from air, can be produced only by intensive action, for instance, by a ship's screws turning in water or in the course of propagation of ultrasonic waves (see Section 28-8). Such cavities cannot exist for a long time inside the liquid. They suddenly changing their shape to spherical, break up and subsequently collapse. The process of generation, motion, transformation and collapse of cavities is termed *cavitation* (from the Latin *cavus* for hollow). It is a major cause of rapid wear of ships' screws.

Thus, liquids have many properties in common with those of solids. However, the higher the temperature of a liquid the closer are its properties to those of dense gases and the more remote from the properties of solids. This means that the liquid state is an intermediate state between the solid and the gaseous state.

Note in addition that the transition of a substance from the solid to the liquid state is accompanied by a less drastic change in properties than the transition from the liquid to the gaseous state. This generally means that the properties of the liquid state of matter are closer to those of the solid state than to those of the gaseous state.

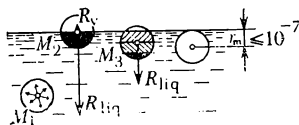
## 12-2 The Surface Layer of a Liquid

Let us find the difference between the action of molecular forces inside the liquid and on its surface.

The average value of the resultant of molecular forces acting on the molecule  $M_1$  inside the liquid is close to zero (Fig. 12.1). The only effect of the fluctuations of this resultant is to produce random motion of the molecule  $M_1$  inside the liquid. The situation is quite different in case of the molecules  $M_2$  and  $M_3$  in the surface layer of the liquid.

Each molecule has a sphere of interaction with a radius of about  $10^{-7} \text{ m}$  (or less). Molecule  $M_2$  then has many more molecules inside its lower hemisphere than inside its upper hemisphere because there is liquid below and vapour and air

Fig. 12.1 All molecules inside surface layer  $10^{-7} \text{ m}$  thick are pulled into liquid.



above. The result is that the net force of molecular attraction acting on the molecule  $M_2$  in the lower hemisphere,  $R_{\text{liq}}$ , is much greater than the net force of molecular interaction in the upper hemisphere,  $R_v$ .

Note that  $R_v$  is so small that it can be neglected. The net force of attraction acting on the molecule  $M_3$  is less than that acting on the molecule  $M_2$  because it is due only to the action of molecules inside the blackened area. It is essential that the net forces acting on the molecules  $M_2$  and  $M_3$  are directed into the liquid at right angles to its surface. Hence, all molecules inside a surface layer with a thickness of the order of the radius of molecular interaction (Fig. 12.1) are pulled into the liquid. But the space inside the liquid is occupied by other molecules; therefore the surface layer sets up a pressure on the liquid termed *molecular pressure*.

If an object is immersed in a liquid, it is surrounded by a layer of liquid in which the molecular forces act in the direction from the object to the liquid, that is, they compress the liquid and do not act on the object.

Molecular pressure manifests itself in such phenomena as compressibility, surface tension, capillarity, etc. Theoretical estimates have shown it to be very great. For instance, for water it is of the order of  $11 \times 10^8 \text{ N/m}^2$  (11 000 atm) and for ethyl ether  $1.4 \times 10^8 \text{ N/m}^2$ .

It is now understandable why it is difficult to compress a liquid by means of external pressure. Indeed, to this end one should set up a pressure of the order of molecular pressure of the liquid. Because this pressure is very high, the task is a difficult one. Hence, if the pressures are not too high, a liquid may be considered incompressible.

### 12-3 Surface Tension

Since the molecules in the surface layer of a liquid are pulled into the liquid, their potential energy exceeds that of the molecules inside the liquid. The same conclusion can be drawn if one takes account of the facts that the potential energy of molecular interaction is negative (see Section 2-5) and that the molecules in the surface layer ( $M_2$  and  $M_3$  in Fig. 12.1) interact with a smaller number of molecules than the molecules inside the liquid ( $M_1$ ).

This additional potential energy of the molecules of the liquid in its surface layer is termed *free energy*, since it may be used to perform work if the free surface of the liquid is reduced. And vice versa, to withdraw a molecule from the inner layers into the surface layer the action of molecular

forces should be overcome, that is, work required to increase the free energy of the surface layer should be performed. It may easily be seen that in this case the variation in free energy,  $\Delta U$ , should be proportional to the variation in the free surface of the liquid,  $\Delta A$ :

$$\Delta U = \sigma \Delta A \quad (12.1)$$

But since  $\Delta U = W$ , we have

$$W = \sigma \Delta A \quad (12.2)$$

Hence, the work of molecular forces,  $W$ , performed in reducing the free surface of a liquid is directly proportional to  $\Delta A$ . But this work should also depend on the nature of the liquid and on external conditions, for instance, temperature. This dependence is expressed by the coefficient  $\sigma$ .

The quantity  $\sigma$  characterizing the dependence of the work of molecular forces performed in the process of changing the free surface area of a liquid on the nature of the liquid and on the external conditions is termed *coefficient of surface tension* of the liquid or simply *surface tension*. In other words, surface tension is the work performed by molecular forces when the free surface area of the liquid is reduced by a unit:

$$\sigma = W/\Delta A \quad (12.2a)$$

Let us derive a unit for measuring  $\sigma$ :

$$\sigma = 1 \text{ J/1 m}^2 = 1 \text{ J/m}^2$$

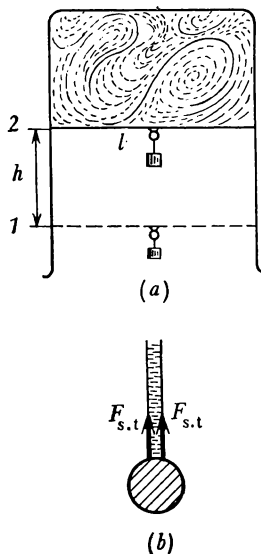
In the SI system the unit of  $\sigma$  is the surface tension of such a liquid the molecular forces of which perform work equal to 1 J when the free surface area of the liquid is reduced by 1 m<sup>2</sup>.

Since all systems spontaneously effect transition into the state with the minimum potential energy, a liquid should spontaneously effect the transition into the state in which the area of its free surface would be minimum. The following experiment proves it.

A wire bent in the shape of  $\Pi$  is provided with a sliding crossbar  $l$  (Fig. 12.2). The loop thus made is covered with a soap film by immersing it into a soap solution. After the loop is withdrawn from the solution, the crossbar  $l$  moves upwards, that is, the molecular forces reduce the free surface area of the liquid. (What happens to the energy liberated in the process?)

Since the minimum surface corresponding to a given volume is that of a sphere, a liquid in weightlessness assumes the shape of a sphere. The same is the reason for the spherical

Fig. 12.2 (a) Molecular forces reduce free surface area of liquid, displacing crossbar from position 1 to position 2; (b) cross section of the film.



shape of liquid drops. The shape of soap films on different frames always corresponds to the minimum free surface area of the liquid.

#### 12-4 Measuring Surface Tension

The molecule  $M_1$  at the surface of a liquid (Fig. 12.3) interacts not only with the molecules inside the liquid but also with the molecules on its surface that are within the sphere of molecular interaction. The net force of molecular interaction,  $R$ , lying in the plane of the surface and acting on the molecule  $M_1$  is zero, but for the molecule  $M_2$  at the edge of the surface  $R$  is not zero. It may be seen from Fig. 12.3 that the direction of the force  $R$  is normal to the boundary of the free surface and tangent to the surface itself.

Every element of a closed line lying on the surface of a liquid is acted upon by molecular forces normal to it and lying on the surface so as to contract the area of the liquid's surface bounded by the closed line. This may be demonstrated with the aid of the following experiment.

A thread of length  $l$  is fastened to a wire ring (Fig. 12.4a). If the entire area of the ring is covered with a soap film, the thread will lie freely on the film's surface, since molecular forces will try to contract both the lower and the upper surfaces bounded by the thread. If we break the film below the thread, the molecular forces will contract the upper surface, straining the thread (Fig. 12.4b). Since the molecular forces act on all the molecules along the thread, the total force acting on the thread should be proportional to its length,  $l$ .

The force  $F_{s,t}$  due to the interaction of the molecules of the liquid aimed at contracting the free surface area of the liquid and tangent to this surface is termed the *surface tension force*. The surface tension force acting on the crossbar  $l$  is shown in Fig. 12.2a. Let us demonstrate now that  $F_{s,t}$  is proportional to  $l$ .

The work performed by the force  $F_{s,t}$  in displacing the bar  $l$  from position 1 to position 2 by the distance  $h$  is expressed by formula (12.2):  $W = \sigma \Delta A$ .

On the other hand, the work  $W$  may be found if one multiplies the force by the path. Since in our case the film contacts the bar along two lines (Fig. 12.2b), the total force in our example is  $2F_{s,t}$ , or  $W = 2F_{s,t}h$ . The reduction in the area of the free surface of the liquid  $\Delta A$  corresponding to the displacement  $h$  of the bar is  $2hl$ ; therefore

$$2F_{s,t}h = \sigma 2hl, \quad \text{or} \quad F_{s,t} = \sigma l \quad (12.3)$$

Fig. 12.3 In addition to forces drawing molecule  $M_2$  into liquid,  $M_2$  is acted upon by forces directed along liquid's surface.

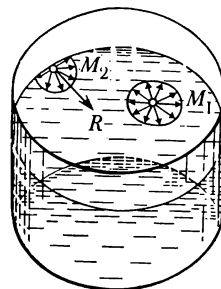
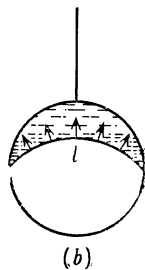
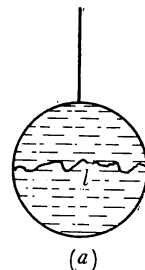


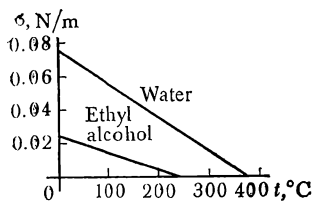
Fig. 12.4 (a) Thread  $l$  lies freely on film; (b) molecular forces acting along liquid's surface (indicated by arrows) pull thread.





**Table 12.1** Surface tension of water and ethyl alcohol

$t, ^\circ\text{C}$	$\sigma, \text{N/m}$	
	Water	Ethyl alcohol
0	0.0756	0.0249
30	0.0712	0.0219
60	0.0662	0.0192
90	0.0608	0.0164
120	0.0549	0.0134
150	0.0486	0.0101
180	0.0423	0.0067
210	0.0354	0.0033
240	0.0286	0.0001
243		0
300	0.0144	
370	0.0005	
374	0	

**Fig. 12.5** Surface tension versus temperature for water and ethyl alcohol.

Hence

$$\sigma = F_{s,t}/l \quad (12.3a)$$

It follows from (12.3a) that surface tension is numerically equal to the surface tension force acting on unit length of the boundary of free surface of the liquid.

Recall that the unit for measuring  $\sigma$  is  $1 \text{ J/m}^2$  and

$$1 \text{ J/m}^2 = 1 \text{ N}\cdot\text{m/m}^2 = 1 \text{ N/m}$$

this following directly from formula (12.3a).

Experiments show that other factors affecting the magnitude of surface tension are the medium in contact with the liquid's surface and the temperature of the liquid. With the rise in temperature the surface tension of the liquid decreases (explain why) and drops to zero at the critical point (see Table 12.1 and Fig. 12.5). This is another proof of the fact that at the critical point a liquid and its vapour become undistinguishable.

Now it may be easily understood why the liquid assumes the shape for which the area of its free surface becomes minimum: the forces of molecular interaction draw the molecules from the surface into the liquid and the forces of surface tension reduce the area of the free surface, that is, close the "gaps" in this surface produced by the withdrawal of surface molecules.

Hence, the surface layer of a liquid is always in a state of stress. This stressed state cannot, however, be likened to the stress in an extended elastic film. Elastic forces increase as the surface is extended, whereas the forces of surface tension are independent of the surface area. The force  $F_{s,t}$  in position 1 and 2 (Fig. 12.2) is the same. The reason for the difference is that in the extended elastic film the number of molecules per unit area decreases as the film is extended; the number of molecules per unit area of the free surface of a liquid, on the other hand, remains constant no matter what the surface's area is.

## 12-5 Wetting

If a glass rod is immersed in mercury and then withdrawn, there will be no mercury on it. If the same rod is immersed in water, there will be a drop of water on its end after it is withdrawn. This experiment shows that the mutual attraction of the mercury molecules exceeds their attraction to the glass molecules, the opposite being true for the interaction of the water molecules with glass.

If the mutual attraction of the molecules of a liquid is less than their attraction to the molecules of a solid, the liquid is said to wet the solid. (For instance, water wets clean glass, but does not wet paraffin.) If the mutual attraction of the molecules of a liquid exceeds their attraction to the molecules of a solid, the liquid is said to be *nonwetting* with respect to the solid. Mercury does not wet glass, but it wets clean copper and zinc.

Place a flat plate horizontally and put a drop of the liquid being studied on it. The shape of the drop will be either as shown in Fig. 12.6*a* or as in Fig. 12.6*b*. In the first case the liquid wets the plate and in the second it does not. The angle  $\theta$  in Fig. 12.6 is called the *angle of contact*. This is the angle between the plane surface of a solid and the plane tangent to the liquid's surface and passing through the point *A* of contact of the solid, the liquid and the gas (Fig. 12.6); there is always liquid inside the contact angle. If the liquid wets the solid, the contact angle is acute, and if it does not, the angle is obtuse. To avoid distortion of the contact angle by gravitation it is advisable to use as small drops as possible.

The usual measure of wetting is the cosine of the contact angle,  $\cos \theta$ , which is positive for wetting and negative for nonwetting liquids. Since the contact angle  $\theta$  retains its value when the solid surface is placed in the vertical position (Fig. 12.7*a*), a liquid which wets the walls of the vessel rises at the walls, and a liquid which does not wet the walls sinks at the walls (Fig. 12.7*b*).

For perfect wetting,  $\cos \theta = 1$ . In this case the liquid spreads over the entire surface of the solid. In the case of perfect wetting it is impossible to obtain a drop on a horizontal surface of the solid. For instance, water perfectly wets a clean surface of glass. Note that for a perfectly nonwetting liquid,  $\cos \theta = -1$ . A small drop of a perfectly nonwetting liquid on a horizontal surface assumes the shape of a sphere.

## 12-6 The Shape of Liquid Surfaces

The fact that the surface of a liquid curves at the vessel's walls may be easily observed in experiment. This is especially visible in narrow tubes where the entire surface of the liquid is curved. In a tube of circular cross section this surface constitutes a part of a sphere and is termed *meniscus*. A wetting liquid forms a concave meniscus and a nonwetting a convex one (Fig. 12.8).

Fig. 12.6 Contact angles  $\theta$  acute for wetting liquid (a), and obtuse for nonwetting (b).

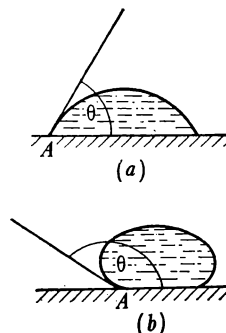


Fig. 12.7 Shape of liquid surface near vessel's wall: (a) wetting liquid; (b) nonwetting liquid.

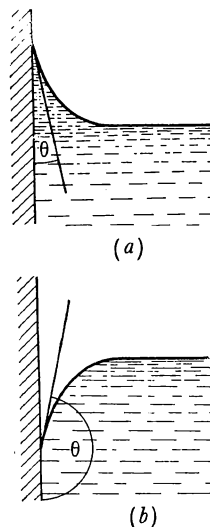
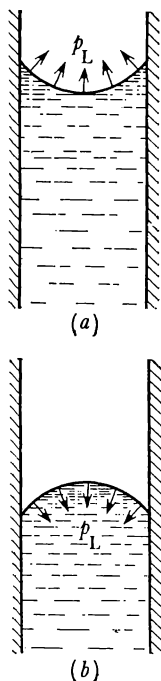


Fig. 12.8 In narrow cylindrical tube, surface of wetting liquid forms concave meniscus (a); that of nonwetting liquid, convex meniscus (b).



Since the meniscus surface greatly exceeds the tube's cross section, the curved surface of the liquid acted upon by molecular forces tries to flatten out and thereby there appears an additional pressure  $p_L$  which in the case of wetting (a concave meniscus) is directed downwards and in the case of nonwetting (a convex meniscus) is directed upwards. The magnitude of this pressure was first calculated by the French mathematician Pierre S. Laplace (1749-1827) and because of that it is often called the *Laplace pressure*.

The expression for this pressure in the case of a spherical free surface of the liquid of radius  $R$  is

$$p_L = 2\sigma/R \quad (12.4)$$

For a surface with an arbitrary shape the Laplace pressure is expressed by the formula

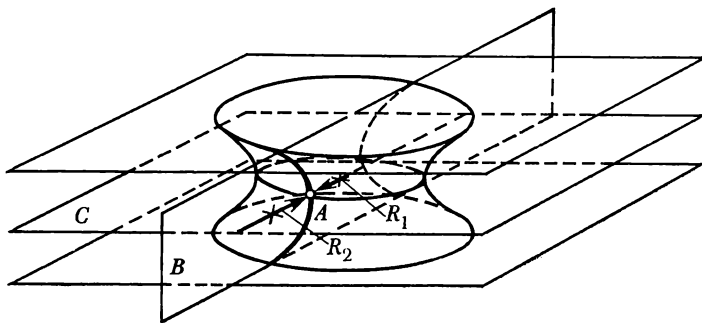
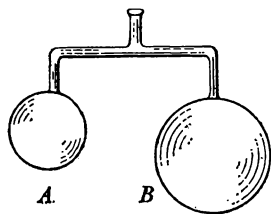
$$p_L = \sigma \left( \frac{1}{R_1} + \frac{1}{R_2} \right) \quad (12.5)$$

Here  $R_1$  and  $R_2$  are the radii of curvature of the surface of the liquid at point  $A$  (Fig. 12.9) defined as the radii of curvature of the lines of intersection of the surface with two mutually perpendicular planes  $B$  and  $C$  (in Fig. 12.9 plane  $B$  is vertical and plane  $C$  horizontal). For a convex line  $R$  is positive and for a concave negative. In our example  $R_1$  should be substituted into formula (12.5) with a plus sign and  $R_2$  with a minus.

The dependence of the Laplace pressure on the radius of curvature of a liquid's surface may be demonstrated in the following experiment. With the aid of a special tube two soap

Fig. 12.9 Drop of wetting liquid between two plates (planes  $B$  and  $C$  are at right angles).

Fig. 12.10 Laplace pressure drives air out of bubble  $A$  into bubble  $B$ .



bubbles of different sizes are blown (Fig. 12.10). If the tube's inlet is closed, the small bubble will be seen to diminish in size and the big one to grow. The explanation is that the

Laplace pressure in the small bubble exceeds that in the big one and it forces the air out of the small bubble into the big one.

## 12-7 Capillarity

The formation of curved liquid surfaces in narrow tubes results in the apparent violation of the law of communicating vessels. If a narrow glass tube is immersed in water (Fig. 12.11a), the latter will be drawn into the tube and its level will be higher than outside. The explanation is that the Laplace pressure in the tube is directed upwards. It draws water into the tube until it is compensated by the hydrostatic pressure  $p_h$  of the column of water in the tube, that is,  $p_L = p_h$ . Since  $p_L = 2\sigma/r$  (in case of perfect wetting the radius of the spherical surface  $R$  in (12.4) is equal to the internal radius of the tube  $r$ ) and  $p_h = \rho gh$ , we have  $2\sigma/r = \rho gh$ , or

$$h = \frac{2\sigma}{\rho gr} \quad (12.6)$$

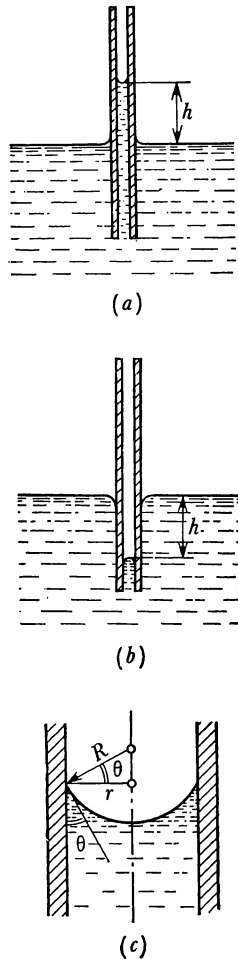
It is evident now that  $h$  is the greater the smaller is the internal radius of the tube  $r$ . Water rises to a substantial height in tubes whose inner diameter is comparable to the thickness of human hair (or even less). Because of that such tubes are termed *capillaries* (from the Latin *capillaris* for hair-thin). A wetting liquid rises in a capillary and a nonwetting sinks (Fig. 12.11a and b). Phenomena resulting from the wetting liquids being drawn into the capillaries and from nonwetting liquids being pushed out of them are termed *capillary phenomena*.

Formula (12.6) makes it possible to estimate the height  $h$  for a zero contact angle  $\theta$ . If the contact angle is nonzero (Fig. 12.11c), the formula for the rise of a wetting liquid or for the drop of a nonwetting liquid in a capillary of radius  $r$  is

$$h = \frac{2\sigma}{\rho gr} \cos \theta \quad (12.7)$$

Suppose two plane glass plates a short distance  $2r$  away from each other are immersed into water. In that case the shape of the water surface between them will be cylindrical (Fig. 12.12a). If two mutually perpendicular and intersecting planes  $B$  and  $C$  are drawn through point  $A$ , the radius of curvature  $R_1$  of the line of intersection in plane  $B$  will be  $r$  (Fig. 12.12b), and the line of intersection of the water surface with plane  $C$  will be straight, that is  $R_2 = \infty$ . Hence, in

Fig. 12.11 Wetting liquid rises in narrow tube (a), nonwetting one sinks in it (b); in case of imperfect wetting ( $\theta \neq 0$ ) meniscus radius is  $R = r/\cos \theta$  (c).



this case the Laplace pressure will be

$$p_{\text{L}} = \sigma \left( \frac{1}{r} + \frac{1}{\infty} \right) \quad \text{or} \quad p_{\text{L}} = \frac{\sigma}{r} \quad (12.8)$$

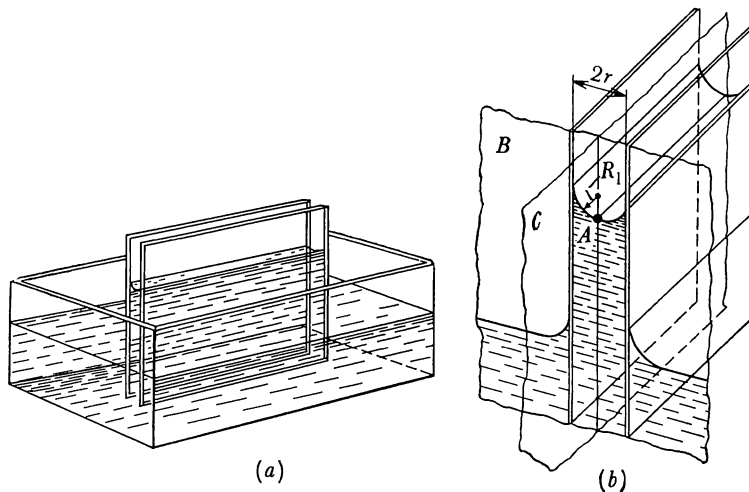
Therefore water will rise between the plates by an amount

$$h = \frac{\sigma}{\rho g r} \quad (12.9)$$

(Why is Laplace pressure in a soap solution half of that in a soap bubble of equal diameter?)

The level of a liquid in a capillary depends on temperature. The higher the temperature the lower  $\sigma$  and, therefore,  $h$

**Fig. 12.12** (a) Surface of water between two close parallel glass plates is cylindrical; (b) cross section of water surface in plane *B* is circular with radius  $R_1$  ( $r$  in perfect wetting) and cross section in plane *C* is straight line.



in formulae (12.6) and (12.7), that is, the level of water in capillaries lowers with the rise in temperature.

Different liquid levels may be observed in communicating vessels of different diameters if one or two of them are small enough (Fig. 12.13).

Capillary phenomena play an important role in nature and technology. There are numerous tiniest capillaries in plants. In trees, water from the soil rises in the capillaries to reach the crowns and to evaporate from the leaves. There are also capillaries in the soil, the narrower the denser the soil. Water rises in these capillaries to the surface and evaporates rapidly, leaving the soil dry. Early spring plowing disrupts the capillaries, that is, helps to retain subsoil water and thereby to increase the crop.

Capillary phenomena play an important role in technology, for instance, in drying of capillary-porous materials.

Capillary phenomena are also important in construction. For instance, to prevent brick walls from absorbing water, a water-tight spacer free from capillaries is inserted between the foundation and the wall of a building. In the paper industry capillarity has to be taken into account when fabricating different brands of paper. For instance, in the fabrication of writing paper it is impregnated with a special solution to close the capillaries. In everyday life capillary phenomena are utilized in wicks, in blotting paper, in nibs to transport ink, etc.

## 12-8 Viscosity

When a boat is in motion, forces appear in water obstructing this motion. Obstructing forces resulting from the motion of a body in a fluid or a gaseous medium are termed *forces of resistance of the medium*.

A very important peculiarity is that those forces do not include static friction (Fig. 12.14).

With the increase in the speed of a body the resistance forces begin to increase rapidly because the body drags away with it particles of the medium and sets layers of the medium in motion relative to each other. For this reason much energy is expended to move a body at high speed in a medium. To reduce the expenditure of energy one has to know the factors responsible for the magnitude of the resistance forces. The air resistance for a cigar-shaped body,  $F_c$ , is 30 times less than for a disk-shaped body,  $F_d$ , and 5 times less than for a ball-shaped body,  $F_b$ , all bodies being of equal cross-sectional area and moving at equal speed  $v$  (Fig. 12.15).

The dependence of the force of resistance of a liquid on the body's speed is fairly complex and is determined both

Fig. 12.13 Levels of liquid in legs of U-tube are different if one tube is capillary: (a) U-tube is filled with wetting liquid; (b) U-tube is filled with nonwetting liquid.

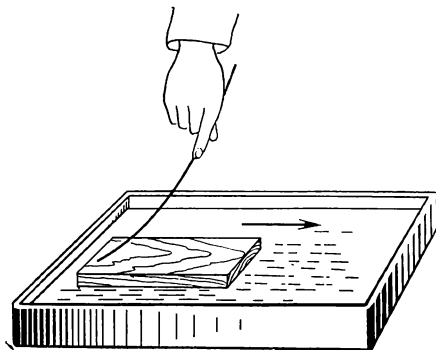
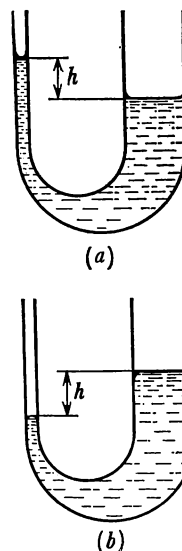
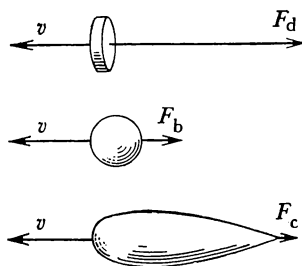


Fig. 12.14 When pressure is applied by means of thin glass filamen to wooden block, its speed slowly rises.

**Fig. 12.15** Influence of shape of body moving in air on air resistance.



by the nature of the motion of parts of the liquid with respect to each other and by the properties of the liquid itself.

The motion of a part of liquid relative to other parts of it creates forces of resistance termed *forces of internal friction*, or *forces of viscosity*. The forces of internal friction tend to make adjacent parts of the liquid move at the same speed.

Motion of a fluid is termed *laminar*, or *streamlined*, if the speed of its motion at any point of the liquid remains constant. In the case of laminar flow of a fluid in a cylindrical pipe the maximum speed is that of the fluid moving along the pipe's axis, the flow velocity close to the walls being zero (Fig. 12.16). The entire fluid is subdivided into cylindrical layers whose speed decreases from the axis towards the walls.

Hence, in the case of laminar flow there is no friction between the walls of the pipe and the liquid because the layer of the liquid close to the walls remains at rest, only the internal friction due to the viscosity of the fluid being active.

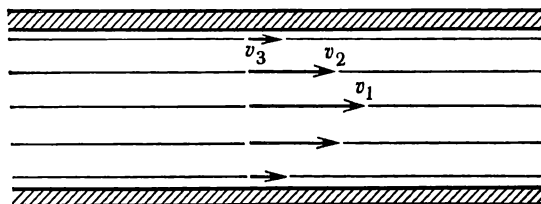
The difference in the velocities of flow of the layers of fluid in the pipe is characterized by the *velocity gradient*. Gradient is a vector characterizing the rate of change of a quantity in space. This vector is tangent to the line along which the rate of variation of the quantity is maximum and points in the direction of its increase. The numerical value of the gradient is the magnitude of the variation of the quantity per unit length of the aforesaid line. The designation for the velocity gradient is  $\text{grad } v$ .

In our example we obtain for radial variation of flow velocity the expression

$$\text{grad } v = \frac{\Delta v}{\Delta r} \quad (12.10)$$

where  $\Delta v$  is the variation of flow velocity in the interval  $\Delta r$  at a distance  $r$  from the axis. Since in actual fact the dependence of  $v$  on  $r$  is nonlinear, formula (12.10) is valid only for very small  $\Delta r$ 's.

**Fig. 12.16** Laminary flow of fluid in cylinder.



## 12-9 Newton's Law of Fluid Friction

Suppose plane layers of a fluid a distance  $\Delta x$  apart move at speeds  $v_1$  and  $v_2$  (Fig. 12.17). In that case the lower layer will accelerate the middle layer and the upper layer will slow it down. Acted upon by forces of internal friction, the middle layer will move at a speed  $v$  greater than  $v_1$  but less than  $v_2$ . If the dependence of the speed of a layer on its position  $x$  is linear, the magnitude of  $\text{grad } v$  will everywhere be the same and equal numerically to  $(v_2 - v_1)/\Delta x$ , or  $\Delta v/\Delta x$ .

Newton demonstrated that the force of internal friction acting on the middle layer is directly proportional to the velocity gradient and the area of the boundary surface of the layer  $\Delta A$ . Mathematically, Newton's law of fluid friction is expressed by the formula

$$f = \eta \frac{\Delta v}{\Delta x} \Delta A \quad (12.11)$$

The quantity  $\eta$  expressing the dependence of the force of internal friction on the nature of the substance and on external conditions is termed *viscosity coefficient* of the substance.

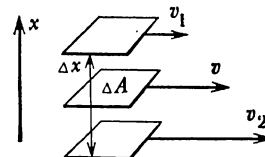
Let us derive a unit for measuring

$$\eta = \frac{f}{\Delta v/\Delta x \Delta A}, \quad \eta = \frac{1 \text{ N}}{\frac{1 \text{ m/s}}{1 \text{ m}} 1 \text{ m}^2} = 1 \frac{\text{N} \cdot \text{s}}{\text{m}^2}$$

In the SI system the unit for measuring  $\eta$  is the viscosity of such a medium in which a force of internal friction of 1 N acts on an 1 m<sup>2</sup> area of a layer at a velocity gradient of 1 s<sup>-1</sup>. The viscosity of a fluid is a function of its temperature. It is noteworthy that the viscosity of gases rises with the temperature and the viscosity of liquids drops. This is an indication that the origins of internal friction of gases and of liquids are different.

The viscosity of gases is due to its molecules flying from one layer to another in the course of their random motion. Since the molecules of a flowing gas in addition to random motion **take part** in the directional motion of the gas flow, the directional velocity being that of the layer with which the molecule moves, the molecules reaching the middle layer from the lower layer which moves at greater speed (see Fig. 12.17) will accelerate its motion and the molecules from the slower moving upper layer will slow it down. Since a rise in temperature is accompanied by an increase in the velocity of random motion of the gas molecules, the trans-

Fig. 12.17 Motion of plane layers of fluid.





port of the molecules from layer to layer is intensified and the viscosity of the gas increases.

The transition of the molecules from layer to layer takes place in liquids as well, but the main cause of a liquid's viscosity is the mutual attraction of its molecules. Because of thermal expansion the forces of intermolecular attraction diminish with the rise in temperature and this explains the decrease in their viscosity. For instance, the viscosity of water at 0 °C is  $17.75 \times 10^{-4}$  N s/m<sup>2</sup> and at 90 °C it is  $3.20 \times 10^{-4}$  N s/m<sup>2</sup>.

The French scientist Jean L. M. Poiseuille (1799-1869) demonstrated in 1840 that the volume of a fluid flowing out of a pipe is, in the case of laminar flow, proportional to the fourth power of the pipe's radius. At present Poiseuille's law is written in the form

$$V = \frac{\pi r^4}{8l\eta} (p_1 - p_2)t \quad (12.12)$$

Here  $V$  is the volume of the fluid flowing out of a pipe of radius  $r$  and of length  $l$  in time  $t$ , the pressure difference between the ends of the pipe being  $\Delta p = p_1 - p_2$ . Formula (12.12) enables the viscosity of various liquids flowing through the same pipe to be compared, and this is the principle of operation of the instrument called *viscometer*.

## 12-10 Amorphous Substances

Some liquids have great viscosity, for instance, glycerine, and honey. However, the viscosity of resin, tar and liquid glass is still greater, and when these substances are cooled it becomes so great that their molecules lose their mobility, that is, their time of residence becomes very great. Such substances cannot be distinguished by their behaviour from solids, that is, they retain their shape and their volume. However, there is only a short-order in the arrangement of their molecules and no long-order. Therefore the structure of those substances is that of liquids, but of very viscous ones (with great times of residence of the molecules).

When liquids with great viscosities are cooled, forces of molecular interaction tend to bring the molecules together and align them, so that there is a tendency towards long-order arrangement. But because of the great viscosity the molecules do not have time to occupy their natural positions of equilibrium and get "stuck" in transient positions. In other words, the effect of molecular forces is not enough to produce long-order arrangement, and so there are bodies that

look like solids but have an arrangement of molecules characteristic of liquids. Such substances having the properties of a solid but devoid of crystalline structure are termed *amorphous*, or *vitreous* since glass is a typical representative.

The similarity of the amorphous substances to liquids is not only in their internal structure. If their temperature is not very low, they exhibit slow fluidity. For instance, tar spreads slowly. Solids of greater density slowly sink in amorphous substances, and of lower density slowly rise.

Physics regards amorphous substances as supercooled liquids which failed to crystallize due to their great viscosity. Amorphous substances may turn into crystals, the process being a very slow one. Some substances exist both in the crystalline and in the amorphous states.

In conclusion we would like to note that amorphous substances gradually thicken upon cooling and gradually become fluid upon heating. It is impossible to discover a clear-cut boundary between the solid and the liquid states of amorphous substances.

## The Solid State

## 13

### 13-1 What Is a Solid?

A substance is usually called solid if it retains its shape and its volume. Those are, however, only external properties characterising the solid state. From the point of view of physics such properties do not make it possible to distinguish between the solid and the liquid states of a substance. Judged

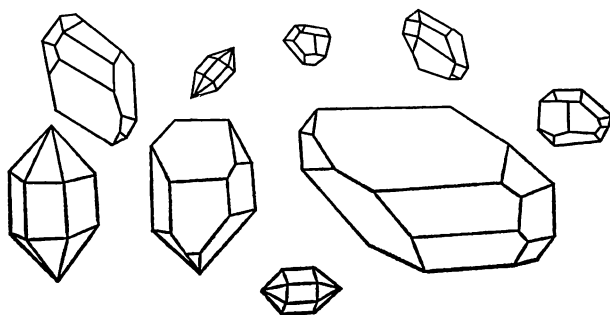


Fig. 13.1 Crystals have definite dihedral angles.

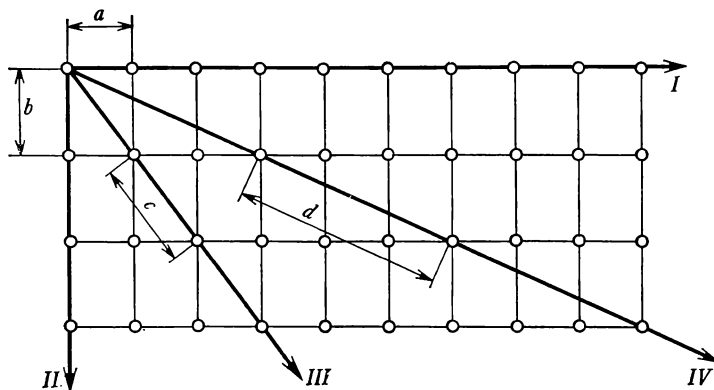
by those properties an amorphous substance would be a solid although its internal structure is that of a liquid.

Many solids were observed to have smooth flat surfaces (faces) arranged at definite angles and sometimes shaped like regular polyhedrons (Fig. 13.1). Such solid bodies are termed *single crystals*. A natural single crystal is usually very small although some single rock crystals are as large as a man's body.

The X-ray studies of internal structure of crystals helped to establish that there is a regular order in the arrangement of the particles making up a crystal (molecules, atoms, ions), that is, that they constitute a *crystal lattice* in space. The points of the crystal lattice corresponding to the stable equilibrium positions of the solid's particles are termed *lattice sites*.

Lattice sites are arranged in a regular order, which is periodically repeated in the crystal. This means that if the distance between nearest sites along a line is  $a$  (Fig. 13.2),

**Fig. 13.2** Distribution of sites along any straight line in crystal lattice is periodic.



the  $n$ th site on the same line will be at a distance of  $na$  from the first site. Such regularity exists throughout the single crystal and is termed *long range order*.

Strictly speaking, a solid in physics is understood to be a crystalline material. It can be made up of numerous small single crystals, and then it is called a *mosaic (imperfect) crystal*, or it can consist of one single crystal, and then it is called a *perfect crystal*. We will elaborate on this difference, and on the difference between amorphous and crystalline material, later on.

### 13-2 Crystalline Anisotropy

The regular arrangement of particles in the crystal lattice is the cause of *crystalline anisotropy*\*, that is, of the dependence of their properties on the orientation of the crystal under test.

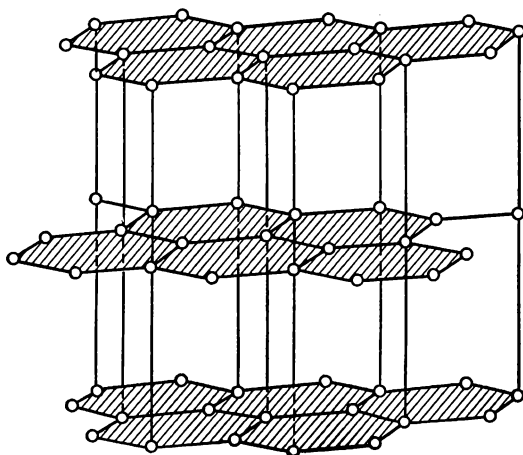
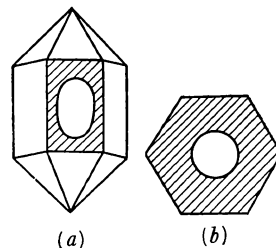


Fig. 13.3 Arrangement of sites in graphite lattice.

The mechanical strength of many crystals is clearly dependent on the direction (i.e. on the orientation of some particular face with respect to the test instrument) in which the force is applied. For instance, mica is easily split into plates, rock salt is easily fractured into cubes, etc. This dependence especially manifests itself in graphite. The atoms in each layer of the graphite crystal occupy positions at the vertexes of regular hexagons (Fig. 13.3), the distance between successive layers being 2.5 times greater than between the atoms in the layer. Because of that the layers of the graphite crystal can be easily shifted with respect to each other. We make use of this phenomena when we write with a pencil. The same property of graphite makes it useful as a lubricant (especially at high temperatures). Note that the term for the surfaces along which it is easiest to cleave a crystal is *cleavage plane*.

If the surface of a quartz crystal is covered with a layer of wax and the end of a hot wire is brought in contact with the centre of its face (Fig. 13.4a), the shape of the molten wax will be that of an ellipse. Such experiments can be used to demonstrate the dependence of numerous other properties of crystals on direction.

Fig. 13.4 The heat conductivity of quartz depends on direction: wax on one face melts and forms ellipse (a); on another face it forms circle (b).



\* Isotropic means having the same properties in all directions and anisotropic means having different properties in different directions.

Note that the properties of a crystal may turn out to be identical in some directions. Cut off the upper half of the crystal shown in Fig. 13.4*a* and repeat the above experiment, bringing the hot point in contact with the centre of the section. In this case the shape of molten wax will be a circle (Fig. 13.4*b*).

It should again be stressed that anisotropy is peculiar only to single crystals. Most solids have a *polycrystalline structure* (from the Greek *polus* for many), that is, they consist of

Fig. 13.5 Crystal lattice of  $\text{CO}_2$ .

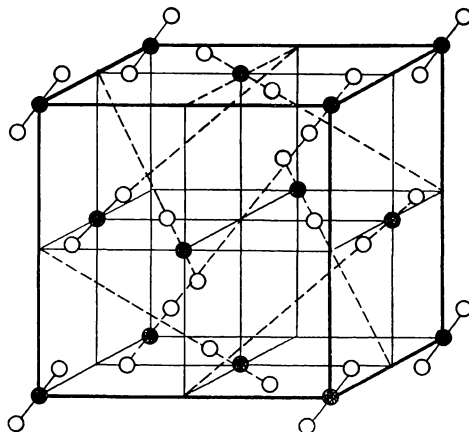
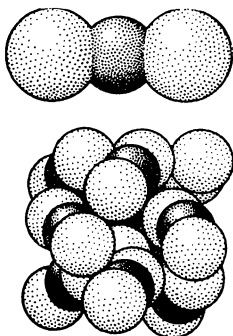


Fig. 13.6 Closely packed arrangement of  $\text{CO}_2$  molecules in carbon dioxide lattice.



many small crystals sometimes visible only in a microscope. Since the crystals are arranged at random, the solid as a whole is isotropic, that is, its properties are the same in all directions (in other words, not depending on the orientation of the specimen with respect to the test instrument) although every constituent crystal is anisotropic. (We note once more that a crystal can have a mosaic structure. The difference between polycrystalline structure and mosaic structure is that the latter consists of single crystals that have grown together so as to be nearly or exactly parallel.) Amorphous bodies are also isotropic because they have no crystal lattice. The difference between amorphous and polycrystalline bodies in this respect is that in most polycrystalline bodies the crystals are large enough to be separated and tested for anisotropy while any part of an amorphous body that can be separated and tested displays no anisotropy.

Figure 13.5 shows the crystal lattice of solid carbon dioxide  $\text{CO}_2$  ("dry ice") with the  $\text{CO}_2$  molecules occupying the sites. Actually the  $\text{CO}_2$  molecules are closely packed inside the crystal like closely packed balls, as shown in Fig. 13.6.

(the  $\text{CO}_2$  molecule is shown above). This is perfectly true of other crystals, as well.

Experiments have shown that the long-range order in the arrangement of the particles of a solid is actually never perfect. The violations of the ideal order in the crystal are termed *defects*, or *imperfections*, of the crystal lattice. One

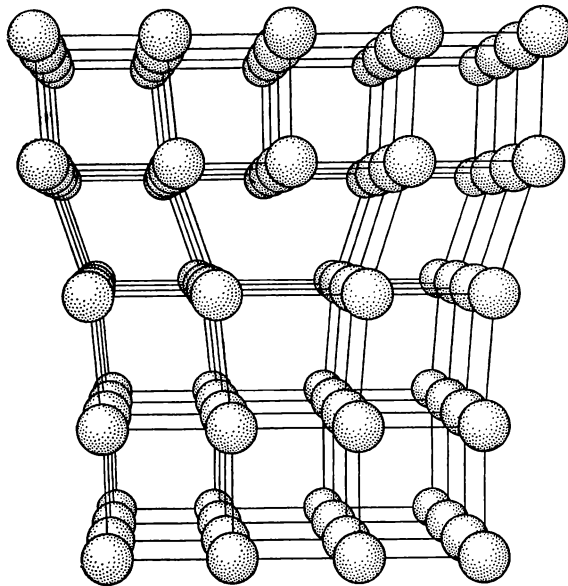


Fig. 13.7 Edge dislocation.

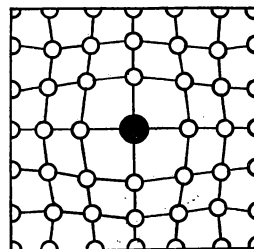
of the most important types of imperfections is the instantaneous displacement of the crystal particles from their equilibrium positions in the course of their thermal motion. Indeed, since the particles are in a constant state of vibration, the sites determine only the average position of each particle.

An important example of a defect is the *edge dislocation* shown in Fig. 13.7.

The more common types of defects are the substitution of a foreign particle for the regular atom (or molecule) in a site (Fig. 13.8), the intrusion of an atom (of the substance itself or foreign) in a position between atoms in regular sites (the defect is called *interstitial*), and the absence of an atom from a regular lattice site (the presence of a *vacancy*). These are not the only types of lattice defects.

The imperfections of the crystal lattice greatly affect many of its properties such as strength, plastic flow, electric conductivity, etc.

Fig. 13.8 Foreign atom in lattice.



### 13-3 Types of Crystals

Fig. 13.9 (a) Crystal lattice of NaCl; (b) closely packed  $\text{Na}^+$  and  $\text{Cl}^-$  ions.

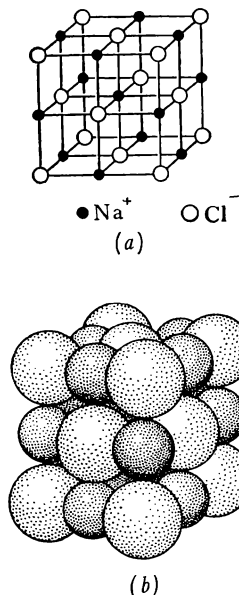
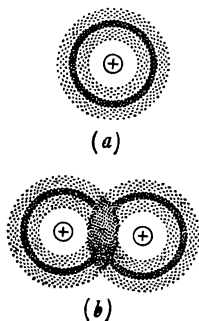


Fig. 13.10 (a) Electron cloud of atom; (b) covalent bond involves overlapping of electron clouds of valence electrons of two atoms.



Various types of crystals and possible arrangement of sites in a crystal lattice are studied in *crystallography*. Physics studies crystal structure not from the point of view of its geometry but from the point of view of forces acting between crystal particles, that is, from the point of view of interparticle bonds. There are four distinct types of crystal structure from the point of view of the forces of interaction: ionic, valence, molecular and metallic. We will consider the essential difference between those structures.

The *ionic* crystal is characterized by the presence of positive and negative ions in the lattice sites. The binding forces holding the ions together in the lattice of this type are electrostatic forces of attraction and repulsion acting between the ions.

Figure 13.9a shows the crystal lattice of NaCl (common salt), where the sites are the centres of the corresponding ions; in Fig. 13.9b the close-packed system of the  $\text{Na}^+$  and  $\text{Cl}^-$  ions in this lattice is shown.

The ions with unlike charges in the lattice are situated closer than the ions with like charges, and because of that the forces of attraction between unlike ions of the lattice prevail over the forces of repulsion between like ions. This explains the substantial strength of ionic crystals.

When solids with an ionic lattice melt, the ions go over into the melt and become free charge carriers. This is the reason for good electric conductivity of such melts. This holds also for aqueous solutions of ionic crystals. For instance, an aqueous solution of common salt is an excellent conductor.

The *valence* crystal is characterized by the presence in lattice sites of neutral atoms with covalent bonds between them. *Covalent* is the term for the bond holding two adjacent atoms together by attraction forces resulting from the atoms exchanging valence electrons, a pair of electrons (one from each atom) taking part in the formation of one bond.

At this point it should be remarked that modern physics is able to calculate the probability of locating an electron in some region of space occupied by the atom. This space may be represented by an electron cloud whose density is proportional to the probability of locating the electron, that is, the cloud is denser in parts more often frequented by the electron (Fig. 13.10a).

The electron clouds of atoms in a covalent molecule overlap. This means that both valence electrons (one per atom) are collectivized, that is, belong simultaneously to both

atoms, spending most of their time between the atoms and connect them into a molecule (Fig. 13.10*b*). This bond is a strong one. The molecules of  $H_2$ ,  $N_2$  as well as molecules made up of different atoms,  $H_2O$ ,  $NH_3$ ,  $SO_2$ ,  $CH_4$ , may serve as examples of molecules with covalent bonds.

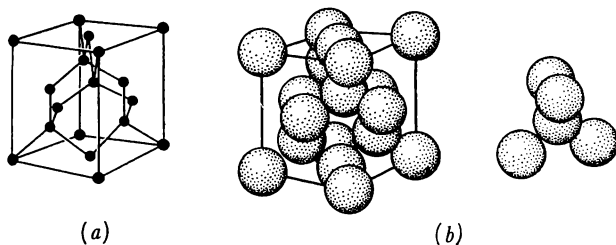


Fig. 13.11 (a) Diamond-type valence crystal lattice; (b) arrangement of atoms (each atom is surrounded by four nearest neighbours).

Solids with a valence crystal structure are quite numerous; for instance diamond, quartz, germanium, silicon. A schematic diagram of the diamond lattice and the packing in it is presented in Fig. 13.11. Each atom of this lattice forms covalent bonds with four neighbouring atoms. Germanium and silicon also have lattices of the diamond type. Covalent bonds produce crystals of exceptional strength. Because of that such crystals have great mechanical strength and high fusion temperatures.

A feature of the *molecular* crystal is the presence of neutral molecules in its lattice (Figs. 13.5 and 13.6) although the same type of bond (termed the *van der Waals bond*) holds together some atomic crystals (for instance, solidified inert gases). This bond is the result of forces acting between neutral

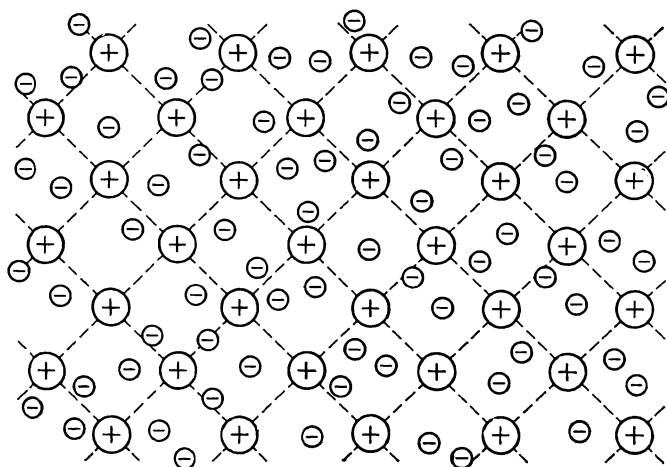


Fig. 13.12 Crystal lattice circles with pluses denote positive metallic ions in lattice sites, and circles with minuses denote electrons.



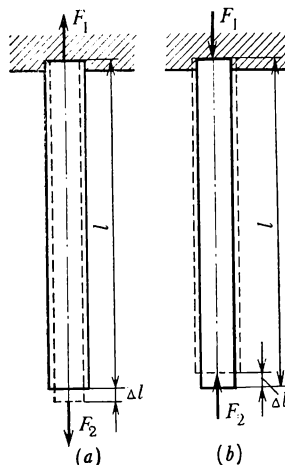
atoms or molecules spaced at distances exceeding those at which the exchange of electrons is possible. Those forces are weak. The solids with the molecular lattice are easily destroyed by mechanical action and have low melting points. Examples of solids with molecular lattice are naphthalene, solid nitrogen and most organic substances.

A feature of the *metallic* crystal is the presence in its sites of positive ions of a metal. The valence electrons (the electrons most distant from the nucleus) of all metals are loosely bound to the ion. The electron clouds of such outer electrons embrace many ions of the metal's crystal lattice. This means that the valence electrons in the crystal lattice cannot belong to one or even to two ions but are collectivized by many atoms. Such electrons can move practically unhampered between the atoms (Fig. 13.12).

Hence, in a solid metal all the atoms lose their outer electrons and turn into positive ions, the electrons moving freely throughout the space occupied by the crystal. This "sea" of electrons holds the positive ions in the sites of the lattice and provides for the great strength of metals.

As the first approximation, the motion of the free electrons in a metal may be treated as the motion of molecules of an ideal gas. For this reason the "sea" of free electrons in a metal is sometimes termed *electron gas* and in calculations the formulae derived for the ideal gas are applied to it. (Using this method, calculate the average velocity of thermal motion of electrons in a metal at  $0^\circ\text{C}$ .) The presence of the electron gas in metals explains the high values of heat and electric conductivities of all metals.

Fig. 13.13. Deformations: (a) longitudinal extension; (b) longitudinal contraction.



### 13-4 Types of Deformation

The variation of the shape or the volume of a body is termed *deformation*. Let us discuss the deformations resulting in practice from the action of mechanical forces on solids.

If the forces  $F_1$  and  $F_2$  are applied to the butts of a rod in opposite directions along its axis, the rod will either extend or contract (Fig. 13.13). The elongation of a body acted upon by forces extending it in one direction is termed *longitudinal extension* (Fig. 13.13a). The decrease in the length of a body acted upon by forces compressing it in one direction is termed *deformation of longitudinal contraction* (Fig. 13.13b). Note that such deformations are accompanied by some variation in the cross-sectional area of the body.

If we pump air into a rubber tube, we see it expand in all directions. If we submerge an inflated tube into water, we

see it contract in all directions (Fig. 13.14). The expansion of the volume of a body acted upon by forces that extend it in all directions is termed *volumetric extension*. The decrease in the volume of the body acted upon by forces that compress it in all directions is termed *volumetric contraction*.

If one end of a horizontal rod is fixed and a vertical downward force  $F$  is applied to its free end (Fig. 13.15a), the rod will bend. Place the rod on two supports and apply to its middle a force  $F$  normal to it (Fig. 13.15b). The rod will sag. The bending of a rod acted upon by forces perpendicular to its axis is termed *lateral bending*. The distance between points  $O$  and  $O'$  in Fig. 13.15b is termed *sag*. Note that in the course of bending the convex side of the rod is extended and the concave contracted.

If both ends of a steel ruler are pressed together it will buckle (Fig. 13.16). The bending of a rod in the course of longitudinal compression is termed *buckling*.

Let us fix a block and apply a force  $F_1$  to shift it (Fig. 13.17). A force  $F_2$  equal in magnitude and opposite in direction will appear at the place where the block is fixed. The action of these forces will cause the block to warp by an angle  $\theta$ , with the upper layers of the block shifting relative to the lower layers (as the pages of a book). The deformation resulting from relative shifting of parallel layers of a body acted upon by forces parallel to these layers is termed *shear*.

Applying two force couples to the butts of a rod so as to turn them in opposite directions (Fig. 13.18), we can observe the rod to twist. In twisting layers of the rod parallel to the butts rotate by an angle. This is evident from the bending of the generatrix  $AB$ . The deformation caused by the relative rotation of parallel layers of a body acted upon by two force couples is termed *twisting*.

Each of the deformations described above can be greater or less in magnitude. A possible measure for each is the *absolute deformation*  $\Delta a$ . Absolute deformation is the variation of the numerical value of some dimension of a body acted upon by forces. For instance, the absolute deformation for longitudinal extension (contraction) of a body is the variation of the body's length  $\Delta l$  (Fig. 13.13), for volumetric extension (contraction) it is the variation of the volume  $\Delta V$ , etc.

However, the relative deformation  $\varepsilon$  (the Greek letter *epsilon*) is a more informative criterion of the variation of the volume or shape of a body acted upon by forces. The relative deformation is the ratio of the absolute deformation  $\Delta a$  to the initial dimension  $a$  of the body:

$$\varepsilon = \Delta a / a \quad (13.1)$$

Fig. 13.14 Volumetric contraction.

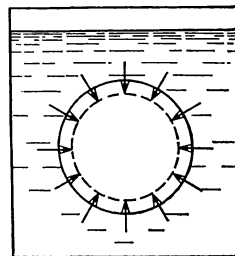


Fig. 13.15 Lateral bending.

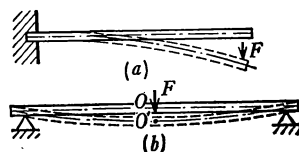
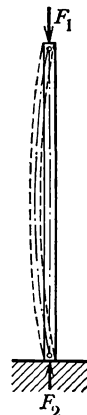
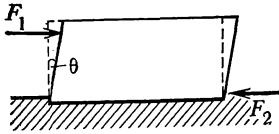


Fig. 13.16 Buckling.



**Fig. 13.17** Shear: magnitude is characterized by shear angle  $\theta$ .



For instance, in the case of longitudinal extension (contraction) we have

$$\varepsilon = \Delta l / l \quad (13.2)$$

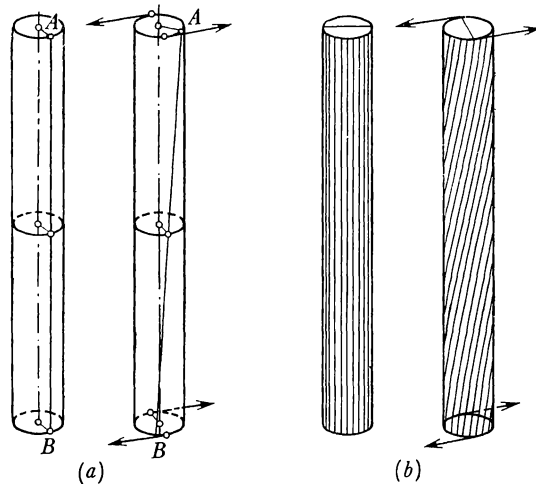
The measure of relative shear is  $\tan \theta$  (why?):

$$\varepsilon = \tan \theta \quad (13.3)$$

### 13-5 Stress

In a deformed solid the displacement of particles of the crystal lattice from their equilibrium positions creates internal forces. External forces applied to a body are transmitted by its lattice to the internal parts of the body and lead to

**Fig. 13.18** (a) Twisting results in applying force couple; (b) twisting curves the cylinder and its generatrices and turns them into helices.



the appearance of stresses and corresponding deformations throughout the body. The presence of internal forces in a strained body can be demonstrated with the aid of the following example.

Imagine a thin layer perpendicular to the axis of the strained rod shown in Fig. 13.13 (Fig. 13.19). It will divide the rod in two parts. Since all parts of the rod will be in a state of equilibrium, the upper part will act on the slice with a force  $F'_1$  equal to  $F_1$  (the weight of the rod is neglected) and the lower with the force  $F'_2$  equal to  $F_2$ . The forces appearing inside a strained body are termed *internal forces*. They cause the deformation of an element inside the body (in our case, an extension).

In a uniform rod acted upon by external forces  $F_1$  and  $F_2$  directed along its axis, the internal forces  $F'_1$  and  $F'_2$  are uniformly distributed over the cross section  $A$ .

The quantity characterizing the action of internal forces in a strained solid is termed *stress*. The measure of stress,  $\sigma$ , is the internal force acting on a unit area of the strained body's cross section:

$$\sigma = F/A \quad (13.4)$$

We derive a unit for measuring  $\sigma$ :

$$\sigma = F/A, \quad \sigma = 1 \text{ N}/1 \text{ m}^2 = 1 \text{ N/m}^2 = 1 \text{ Pa}$$

The unit of  $\sigma$  in the SI system is the stress caused by an internal force of 1 N that acts on a cross section of 1 m<sup>2</sup>.

If the distribution of internal force over the cross section is nonuniform, in (13.4) instead of  $A$  one should use an area  $\Delta A$  sufficiently small for the internal force acting on it to be considered constant.

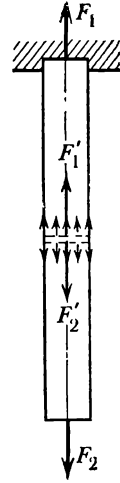
If the internal force acts at right angles to the cross section, the stress is termed *normal stress*  $\sigma_n$  (for instance, in case of longitudinal extension). If this force acts parallel to the cross section, the stress is termed *shearing stress*  $\sigma_{sh}$  (for instance, as in Fig. 13.17).

Stress acting at any point of a strained body can always be represented as the result of action of stresses of those two types, since a force acting on a specific area  $\Delta A$  can be resolved into two components: one perpendicular to the area and the other tangent to it.

### 13-6 Elasticity, Plasticity, Brittleness and Hardness

All types of deformation in a solid result in the relative displacement of particles making up the body. Such displacements create forces opposing the deformation. Those forces, termed *elastic forces*, act both between parts of the strained body and between other bodies causing the deformation. They tend to re-establish the initial shape and volume of the body. The property of strained bodies to re-establish their initial shape and volume after the external forces have ceased to act is termed *elasticity*. Deformations which vanish with the lifting of external stresses are termed *elastic*. Since the elastically deformed body tries to re-establish its original shape and volume, it acts on the bodies causing the deformation with a force termed *elastic force*. Internal forces

Fig. 13.19 Internal forces  $F'_1$  and  $F'_2$ .



appearing in the body in the course of elastic deformation are also termed elastic forces.

Experience shows that a body can be so deformed that it will not be able to re-establish its original shape after the external stresses have been lifted. The property of a body to retain its deformation after the external stresses have been lifted is termed *ductility* (or *plasticity*). The residual deformation of the body remaining after the external stresses have been lifted is termed *plastic deformation*.

The elasticity, or ductility, of bodies is determined by their structure; this, in turn, depends on the material and on its processing. The subdivision of materials into ductile and elastic is a matter of convention since every body is at the same time ductile and elastic. For instance, a steel spring may be so extended that it will fail to contract to its former length; on the other hand, a copper spiral displays the properties of a spring if its elongation is small. Experiments show that as the stress on a body is gradually increased, elastic deformation appears first and is followed by plastic deformation.

The properties of materials are also substantially affected by external conditions. For instance, lead ductile under normal conditions becomes elastic at low temperatures, and elastic steel becomes ductile at high temperatures or at ultra-high pressures.

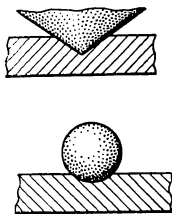
The important mechanical properties of materials that have to be taken into account in engineering are *brittleness* and *hardness*.

In practice one can come across materials which experience elastic deformations at comparatively small stresses but break up as the external load is increased without experiencing plastic deformation. Such materials are termed *brittle* (for instance, glass or brick). Brittle materials are very sensitive to impact stresses. They can be easily broken by a sharp knock.

There are different methods of measuring hardness of a material. Usually the material which leaves scratches on the surface of another is regarded as the harder. Experiment has proved diamond to be the hardest material of all. In engineering practice the hardness of materials is determined by pressing a diamond cone or a steel ball into its surface (Fig. 13.20). The smaller the depth of penetration of the cone into the material (for a constant pressure) the greater the hardness of the material.

The hardness of a material greatly affects the magnitude of rolling friction. For instance, ball bearings are made of hard steel since in this case rolling friction is quite small.

Fig. 13.20 Measuring hardness.



### 13-7 Hooke's Law

The principle of the dynamometer, an instrument for measuring force, is the direct proportionality of the elastic deformation to the force causing this deformation. Conventional spring scales can serve as an example of the aforesaid.

The first to establish the relation between the elastic deformation and the internal forces acting in a material was the British scientist Robert Hooke (1635-1703). The modern version of Hooke's law is as follows: the mechanical stress in an elastically strained body is directly proportional to the relative deformation of this body:

$$\rho = k\varepsilon \quad (13.5)$$

The quantity  $k$  characterizing the dependence of the mechanical stress in a material on its nature and on the external conditions is termed *elasticity modulus*. The measure for the elasticity modulus is the stress which should be established in the body to effect a unit relative elastic deformation.

The unit for measuring the elasticity modulus in the SI system is  $1 \text{ N/m}^2$ . (Prove this!)

Note that **relative** deformation is usually expressed by a number much less than unity. Leaving out few exceptions, it is impossible to obtain  $\varepsilon$  equal to unity since most materials break long before this value is reached. However, the elasticity modulus can be found from small  $\varepsilon$ 's as the ratio  $\sigma/\varepsilon = k$  in formula (13.5) is a constant.

As an example, let us discuss the application of Hooke's law to longitudinal extension, or compression. In this case formula (13.5) assumes the form

$$\sigma = E\Delta l/l \quad (13.6)$$

where  $E$  is the elasticity modulus for this type of deformation known as *Young's modulus*. The measure of Young's modulus is the normal stress which develops in a body whose relative elongation is unity, that is which is extended to twice its length ( $\Delta l = l$ ). Note that the numerical values of Young's moduli are obtained from the results of experiments conducted in the elastic range. In calculations they are usually obtained from tables.

Since  $\sigma_n = F/A$ , we obtain from (13.6):  $F/A = E \Delta l/l$ , whence

$$\Delta l = F/EA \quad (13.7)$$

Hence the absolute deformation in the case of longitudinal extension or contraction is directly proportional to the force acting on the body and to the length of the body and in-

versely proportional to the cross section of the body, being also dependent on the material of the body.

The maximum stress in the material whose lifting re-establishes the original shape and the volume of the body is termed *elastic limit*. Formulae (13.5) and (13.7) are only valid below the elastic limit. When the elastic limit is exceeded, plastic deformations appear in the material. If the stress is increased further, a situation develops when deformation begins to increase without increase in stress and continues even if stress is reduced, which leads to the rupture of the body. The maximum stress which the material can withstand is termed the *breaking point*.

In constructing machinery and buildings the rule is to provide for a safety factor. The *safety factor* is the ratio of the breaking point of the construction material in the most stressed part of the structure to the maximum actual stress.

### 13-8 Energy of a Body Under Elastic Deformation

To strain a body elastically work should be performed. The deformed body gains potential energy  $U$  at the expense of this work and is itself able to perform work  $W$ . For instance, an extended spring attached to a door can close it. Below the elastic limit it can be assumed that  $U = W$ .

It is established that work on the force versus path plot is expressed by the area bounded by the plot and the coordinates of the terminal point. Figure 13.21 shows the dependence of the deformation force,  $F$ , on the deformation,  $\Delta l$ . Suppose this plot applies to the extension of the rod shown in Fig. 13.13a. In that case work  $W$  spent on this extension will be equal to the area of the triangle  $COB$  in Fig. 13.21, that is

$$W = F \Delta l / 2 \quad (13.8)$$

Since for an elastic deformation  $U = W$ , it follows that

$$U = F \Delta l / 2 \quad (13.8a)$$

Since  $F = \sigma_n A$ , we obtain

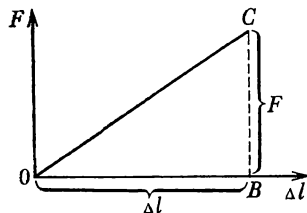
$$U = \sigma_n A \Delta l / 2$$

With account of expression (13.6) we obtain

$$U = \frac{EA}{2l} (\Delta l)^2 \quad (13.9)$$

Hence, the potential energy of a body under elastic deformation is directly proportional to the square of the absolute deformation.

Fig. 13.21 Dependence of force on deformation.



Formula (13.9) may be transformed as follows:

$$U = \frac{EAl}{2} \left( \frac{\Delta l}{l} \right)^2$$

whence

$$\frac{U}{V} = \frac{1}{2} E \epsilon^2 \quad (13.10)$$

where  $V = Al$  is the volume of the body, and  $\epsilon$  is its relative deformation. Formula (13.10) shows that the energy of a body under elastic deformation is distributed over the entire volume of the body. The energy density  $U/V$  of the body is directly proportional to the square of relative deformation and depends on the material.

## Change of State—II

## 14

### 14-1 Fusion and Crystallization

The transition of a substance from the solid state into the liquid state is termed *fusion* (or *melting*) and the transition from the liquid state into the solid is called *solidification* (or *crystallization* if the solid is a crystal).

The fusion of a crystalline substance involves an increase in the distances between the nearest neighbours—atoms or molecules—in the crystal lattice and the destruction of the lattice itself. The average interparticle distances in the melt are usually greater than in the substance, although sometimes the opposite is true (for instance, for many valence crystals). This means that in the process of fusion the potential energy of the particles increases. Hence, fusion cannot take place spontaneously since energy must be spent to effect it.

Upon crystallization the particles (the nearest neighbours) forming the lattice draw closer; that means that their potential energy decreases. Therefore crystallization can take place if the liquid is able to give up its energy to some external body.

Thus, a unit mass of the liquid has greater internal energy than a unit mass of the same substance in the solid state, even if their temperatures are equal. A physically and chemically homogeneous region is termed a *phase*. Note that the solid and liquid phases of the same substance can remain in equilibrium at the same temperature for an indefinite time if they are unable to exchange energy with the surrounding



medium. For instance, ice can float in water for an indefinite time if the temperature of all surrounding bodies will be the same and equal to  $0^{\circ}\text{C}$ .

Suppose there is only the solid phase of a substance and this phase gains energy from other bodies. Then initially the potential and kinetic energies of the particles will rise since both the interparticle distances and the velocities of their motion in the lattice will increase with temperature. Then at a definite temperature the lattice will begin to break up. Until the whole of the substance melts its temperature will remain constant, the entire energy gained being spent on work to overcome interparticle attraction. When only the liquid phase remains, the energy gained will heat the substance, that is, it will raise the kinetic energy of the particles.

If the liquid phase loses energy to the surroundings, all the processes described will recur in the inverse order. Experiments show that for crystals fusion and crystallization take place at a definite temperature, which remains constant as long as the solid and the liquid phases coexist. This temperature is termed *temperature of fusion*. Note that in the two processes there is always a distinct boundary between the liquid and solid phases.

The fusion and solidification of amorphous materials are gradual processes. There is no boundary between the liquid and solid phases, the entire mass of the amorphous substance thickening with the decrease in temperature and thinning with the rise in temperature.

## 14-2 Specific Heat of Fusion

Studies of the processes of fusion and of solidification established the fact that the variation of internal energy in those processes is directly proportional to the mass of the substance  $m$  taking part in them. The variation of internal energy in this case is the heat of fusion  $Q_f$ :

$$Q_f = \lambda m \quad (14.1)$$

The heat of fusion,  $Q_f$ , depends also on the substance and on the surroundings. The quantity  $\lambda$  that characterizes the dependence of the internal energy of a substance in the process of its fusion or solidification on its nature and surroundings is termed *specific heat of fusion*. The measure of the specific heat of fusion is the heat needed to melt a unit mass of the substance at the temperature of fusion:

$$\lambda = Q_f/m \quad (14.1a)$$

We find a unit for measuring  $\lambda$ :

$$\lambda = 1 \text{ J/1 kg} = 1 \text{ J/kg}$$

The unit of  $\lambda$  in the SI system is the specific heat of a substance which requires 1 joule to melt 1 kilogram of its mass at a constant temperature.

The specific heat of fusion is determined from experiments and is taken from tables for calculations. Fig. 14.1 shows

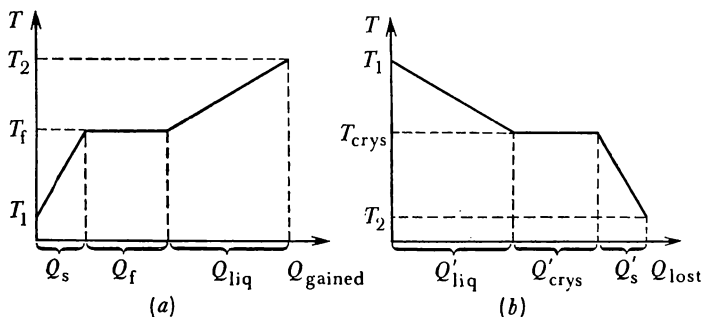


Fig. 14.1 Dependence of temperature of naphthalene on: (a) amount of heat gained; (b) amount of heat lost.

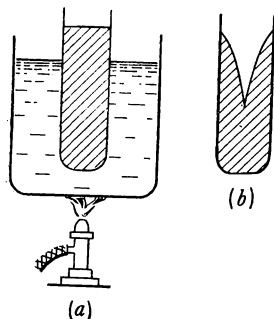
the variation of temperature in the course of fusion and crystallization. The section  $Q_s$  (Fig. 14.1a) is the heat gained by the substance in the solid state in the process of heating,  $Q_f$  is the heat gained in fusion and  $Q_{liq}$  is the heat gained in the process of heating in the liquid state. In Fig. 14.1b,  $Q'_{liq}$  is the heat given up by the liquid in the process of cooling,  $Q'_{crys}$  is the heat given up in crystallization and  $Q'_s$  is the heat given up by the solid in the process of cooling.

### 14-3 Changes in Volume and Density During Fusion and Solidification

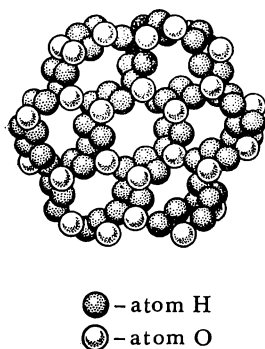
As the result of fusion the regular arrangement of particles in the lattice of a solid changes to an irregular one in the liquid, therefore substantial changes in the volume may be expected to accompany fusion and solidification. Experiments prove this conjecture to be true. For instance, a pocket is formed in molten naphthalene, and when naphthalene is again fused in the test tube, its volume increases substantially (Fig. 14.2).

It was established that the volume of a great majority of substances increases in the process of fusion and decreases in the process of solidification. Obviously, this will lead to changes in density. In the cases mentioned above the density will decrease in the process of fusion and increase in

**Fig. 14.2** (a) Test tube contains molten naphthalene; (b) after cooling and solidification pocket is formed in naphthalene.



**Fig. 14.3** Schematic representation of ice lattice.



the process of solidification. This can be easily proved by experiment. For example, crystals of solid naphthalene sink in molten naphthalene.

Should we make such experiments with bismuth and with water, we would establish that solid bismuth floats in molten bismuth and ice floats in water. Such substances as bismuth, ice, gallium, germanium, silicon contract upon fusion and expand upon solidification. This deviation from the general rule is due to the peculiarities of the crystal lattice of the above-mentioned substances. The crystal lattices of germanium and silicon have low packing densities (they have the diamond-type lattice; see Fig. 13.11). Upon melting the volume of those substances decreases.

Figure 14.3 shows the arrangement of ice molecules in its lattice. It may be seen that the  $\text{H}_2\text{O}$  molecules adhere to one another, but the resulting structure is an open one with substantial pockets in it. Upon melting, the distances between the nearest molecules increase, as in the case of other substances, but the open structure of the crystal is disrupted and the total volume of the substance decreases as the result of molecules filling up the internal pockets. Because of that the density of water turns out to be greater than that of ice.

Investigations have shown that after melting water contains individual parts of the ice lattice which still retain pockets. They break up gradually as the temperature of water is raised. Therefore water contracts upon heating up to the temperature of  $4^\circ\text{C}$ . At  $4^\circ\text{C}$  the destruction of pockets is compensated by the increase in intermolecular distances and water expands upon further heating. When water is cooled, the processes described above are reversed. Hence, there is a single maximum of the density of water at  $4^\circ\text{C}$ .

This property of water is of enormous importance in nature. The expansion of water upon freezing is responsible for the destruction of rock and prevents the complete freezing of lakes and rivers. (Why does the temperature at the bottom of rivers and lakes in winter stay at  $4^\circ\text{C}$ ?)

The variation of the volume of metals in the course of their fusion and solidification is of substantial importance for foundry processes.

#### 14-4 Pressure Dependence of Temperature of Fusion and Heat of Fusion

Experiments show that the variation of external pressure acting on a solid affects its temperature of fusion. In cases when the volume of the substance increases upon melting,

the increase in the external pressure hampers the fusion process and thereby raises the temperature of fusion. If, on the other hand, the volume of the substance increases upon fusion, raising the external pressure facilitates the fusion process and brings about a decline in the temperature of fusion. Note that even great increases in pressure affect the temperature of fusion so little that the effect can practically be neglected. For instance, to decrease the temperature of fusion of ice by 1 K the pressure has to be raised by 130 atm.

The temperature of fusion of a substance at normal atmospheric pressure is termed the *melting point* of the substance.

The specific heat of fusion,  $\lambda$ , has also been found to depend on the pressure. At high pressures the substance in its expansion must perform substantial work against the forces of external pressure. Because of that the specific heat of fusion of substances that expand upon melting increases with external pressure, the opposite being the case for ice, bismuth, gallium, germanium and silicon. For instance, the specific heats at normal pressure are: for mercury  $\lambda_{\text{Hg}} = 11.5 \times 10^3 \text{ J/kg}$  and for bismuth  $\lambda_{\text{Bi}} = 54.5 \times 10^3 \text{ J/kg}$ , the respective values at a pressure of  $12 \times 10^3 \text{ atm}$  being  $\lambda_{\text{Hg}} = 13.2 \times 10^3 \text{ J/kg}$  and  $\lambda_{\text{Bi}} = 38.1 \times 10^3 \text{ J/kg}$ .

#### 14-5 The Law of Heat Exchange for Fusion and Crystallization

Many calculations of heat exchange processes involving fusion and solidification make use of the law of heat exchange. Let us discuss the form this equation takes in the case of measuring specific heat of fusion of ice with the aid of a calorimeter.

Take a calorimeter of mass  $m_{\text{cal}}$  containing a mass of water  $m_{\text{water}}$  at a temperature  $T_1$ . To measure  $\lambda$  of ice, throw a piece of melting ice of mass  $m_{\text{ice}}$  at a temperature  $T_f$  into the calorimeter. Suppose that a temperature  $\Theta$  is established in the calorimeter after the ice has melted completely. In that case it can be assumed that the ice gained heat in the process of melting and the water produced from it gained heat when heated from the melting point to the temperature  $\Theta$ :

$$Q_{\text{gained}} = \lambda m_{\text{ice}} + c_{\text{water}} m_{\text{ice}} (\Theta - T_f)$$

The losers of heat were the calorimeter and the water it initially contained. Therefore

$$Q_{\text{lost}} = c_{\text{cal}} m_{\text{cal}} (T_1 - \Theta) + c_{\text{water}} m_{\text{water}} (T_1 - \Theta)$$

Since  $Q_{\text{gained}} = Q_{\text{lost}}$ , we have

$$\lambda m_{\text{ice}} + c_{\text{water}} m_{\text{ice}} (\Theta - T_f) = (c_{\text{cal}} m_{\text{cal}} + c_{\text{water}} m_{\text{water}}) (T_1 - \Theta)$$

Substituting the results obtained from experiments one is able to calculate the specific heat of fusion of ice. It is equal to  $\lambda_{\text{ice}} = 3.3 \times 10^5 \text{ J/kg}$ .

## 14-6 Solutions and Alloys

It is known from practice that various salts and many other substances (for instance, sugar) easily dissolve in water. It has also been established that in the process of dissolving such substances disintegrate into molecules that form a uniform mixture with the molecules of water. Thus, a *solution* is a homogeneous mixture of the solute and the solvent.

Common salt easily dissolves in water, but one can put so much salt into water that it will no longer dissolve. This is true for most solutions. A solution of a substance which accepts no more of it is termed *saturated*. But there are also solutions in which the substances can mix in any desired proportions; for instance, the solution of ethyl alcohol in water (or of water in ethyl alcohol).

In dissolving solids in liquids energy must be expended; this energy is termed the *heat of solution*. Therefore such dissolution is often accompanied by the cooling of the solution. For instance, when ammonium chloride is dissolved in water, the temperature drops noticeably. Note that when a chemical reaction takes place between the solute and the solvent, the result may be an increase in the temperature of the solution.

Temperature affects the dissolution of many substances. (For instance, the solubility of sugar in water increases with the rise in temperature and that of air in water decreases.) The solubility of many gases drastically increases with pressure. (Recall that at increased pressures large amounts of carbon dioxide can be dissolved in wine or water. A rapid rise from a great depth may be the cause of a diver's death from caisson disease, since a rapid drop in pressure causes gases dissolved in the blood to be discharged and the blood, so to speak, begins to boil.) Solids, liquids and gases can dissolve in liquid solvents. But by no means all substances produce solutions. For instance, mercury and kerosene are not soluble in water.

The crystallization of a saturated solution of a solid may be observed when the solution is either cooled or evap-

orated. This method is expedient for growing large single crystals. To this end a small crystal of the dissolved substance (the seed) is suspended in its saturated solution and the solution is evaporated at a very slow rate.

The solute reduces the temperature of solidification of the solvent and raises its boiling temperature. For instance, a concentrated solution of common salt in water freezes at  $-21^{\circ}\text{C}$  and the solution of calcium chloride at  $-55^{\circ}\text{C}$ .

A mixture of snow and salt is sometimes used as a cooling agent. In such a mixture initially a small amount of solution of salt in water is formed followed by the dissolution in it of the crystals, which results in a substantial decrease in temperature.

Melting different substances and mixing them in definite proportions, we can obtain various *alloys*, for example, metal alloys. Sometimes the result is the formation of a *solid solution*. One example of such is steel. Steel is a solid solution of carbon in iron. The carbon atoms in steel occupy positions in the interstitials of the iron lattice, that is, between the iron atoms.

The atoms of one metal in a solid solution may take the place of atoms of the other. An example of such substitutional solutions of replacement are a copper-gold alloys.

Technology often requires materials with such properties that are not obtainable in natural materials. In such cases new alloys with the necessary properties are developed. Some alloys possess great ductility, others—great mechanical strength and light weight, alloys of the third group have very low temperatures of fusion, those of the fourth are refractories, that is, alloys with very high temperatures of fusion, etc. For this reason the development of new alloys and the study of their properties is one of the most important goals of modern science and technology.

### 14-7 Sublimation

We know from everyday life that many solids have an odour. Examples of such substances are camphor and naphthalene. This proves that under certain conditions substances can turn from the solid state into the gaseous without passing through the liquid state. Indeed, odour is due to the molecules of the solids reaching our nose. This means that the vapours of such substances are contained in the air. All solids in nature evaporate, but in most cases the amount of the vapour is so small that it cannot be detected. As the temperature is raised, the density of the vapour rises rapidly,

which is the reason why many solids begin to give off odours when heated.

The vapourization of a solid is termed *sublimation* (from the Latin *sublimare* for raise). The sublimation of ice or snow may easily be observed. For instance, in winter one may observe hoarfrost to decrease with time.

In food industry solid carbon dioxide  $\text{CO}_2$  ("dry ice") is widely used for refrigeration since it turns directly into gas and thus does not spoil food. The drying of various materials by sublimation is common practice in technology.

The opposite process of direct transformation from gaseous to solid state can also be observed. In winter one can often see ice crystals growing rapidly on window-panes in the shape of beautiful patterns formed directly from water vapour contained in the atmosphere.

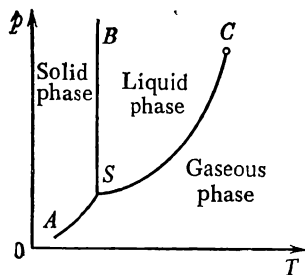
### 14-8 Phase Diagrams. Triple Point

It was mentioned above that the state of a substance depends on the surroundings, mainly on pressure and temperature. This dependence for any substance can be represented on a *phase diagram* in  $p$  and  $T$  coordinates; from this diagram we can easily find what happens to a substance when the state of the surroundings changes.

Figure 14.4 shows this diagram for the case when the *phase space* contains only the substance itself. The curve  $CS$  is the familiar dependence of the saturated vapour pressure of the specific substance on temperature,  $C$  being the critical point and  $S$  corresponding to the temperature of solidification of the liquid in contact with its saturated vapour (when the substance loses energy). The curve  $AS$  expresses the temperature dependence of the pressure of the saturated vapour in equilibrium with the solid's surface. The temperature of fusion depends on pressure, and the line  $BS$  shows this dependence. The part of the diagram to the left of line  $ASB$  corresponds to the solid state of the substance, the part bounded by the line  $BSC$  to the liquid state and, finally, the part of the diagram to the right of the line  $ASC$  to the gaseous state (vapour). The line  $CS$  corresponds to the equilibrium between the liquid and the vapour, the line  $BS$  to the equilibrium between the liquid and the solid and the line  $AS$  to the equilibrium between the solid and the vapour.

Under constant external conditions ( $p$  and  $T$ ) corresponding to some point on the lines of phase equilibrium  $AS$ ,  $BS$  or  $CS$ , two phases of the substance can be in a state of dynamic equilibrium in the course of which equal numbers

Fig. 14.4 The phase diagram.



of molecules go over from one phase to the other, and vice-versa. If energy is not supplied to, or taken away from, the substance, this equilibrium can persist for an indefinite time.

The point  $S$  in the phase diagram corresponding to the equilibrium of all three phases is termed the *triple point*. This point marks the values of  $p$  and  $T$  unique for a specific substance at which the equilibrium of all the three phases of this substance is possible. For instance, at the triple point of water the pressure is 4.58 mmHg and the temperature is 273.16 K (this temperature is used to define the kelvin; see Section 4-4).

When the surrounding medium changes ( $p$ ,  $T$ , or  $p$  and  $T$  simultaneously) the point corresponding to the state of the surroundings moves across the diagram (for instance, heating or cooling at constant pressure correspond to the displacement of the point along a horizontal straight line). When the point on the diagram crosses from one region to another, the state of the substance changes. For instance, if the line  $BS$  is crossed, either fusion or crystallization will take place; if the line  $CS$  is crossed, vapourization or condensation; and if the line  $AS$  is crossed, sublimation or the reverse process. For this reason another term for the equilibrium lines  $BS$ ,  $CS$  and  $AS$  is the *lines of phase transitions* and for the diagram, the *diagram of phase transitions*.

We would like to remind the reader that phase transitions discussed above entail changes in internal energy of the substance, and for their realization require heat of phase transitions: heat of fusion (crystallization), heat of vapourization (condensation), or heat of sublimation to be supplied to, or transported from, the body.

It may be seen from the phase diagram (Fig. 14.4) that sublimation and the reverse process can take place at various temperatures and pressures below those of the triple point. For instance, the pressure and temperature at the triple point of water are (as you know) 4.58 mmHg and 273.16 K; therefore ice can be sublimated only at temperatures below 273.16 K and at pressures below the pressure of saturated water vapour in contact with ice.

The temperature at the triple point of carbon dioxide is  $-56.6^{\circ}\text{C}$  and the pressure is 5.11 atm. Therefore carbon dioxide can exist at the atmospheric pressure only in two states: solid or gaseous, and "dry ice" turns directly into gas; at standard pressure its sublimation temperature is  $-78^{\circ}\text{C}$ .

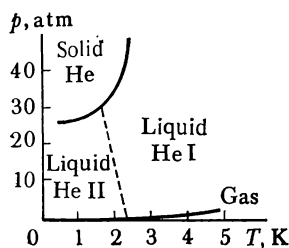
The point  $B$  on the line  $SB$  in most cases is a little displaced to the right from the vertical passing through  $S$ ,



and in the case of ice, bismuth, gallium, germanium and silicon to the left (see Section 14-4). For instance, at point  $S$ , for water,  $p = 4.58$  mmHg and  $T = 273.16$  K (i.e.  $0.01^\circ\text{C}$ ), the melting point for ice (at standard pressure of 760 mmHg) being  $273.15$  K (or  $0^\circ\text{C}$ ).

Note that under transient (nonequilibrium) conditions liquid may exist in the vapour region (superheated liquid) or in the solid region (supercooled liquid). Supersaturated vapour, too, may be found both in the liquid and solid regions. But the solid phase always follows the line  $ASB$  in transitions to the liquid or gaseous states.

**Fig. 14.5** The phase diagram for helium.



A quite peculiar phase diagram is that of helium (Fig. 14.5). It may be seen from this diagram that the curves of equilibrium of the solid phase with the liquid phase and of the liquid phase with the gaseous phase do not intersect in any point, that is, helium has no triple point. Other substances possessing such a property are not known.

The critical temperature of helium is 5.25 K. Consequently, helium can be turned into a liquid only if it is cooled below this temperature. Experiments held by the Soviet scientist Peter L. Kapitza (b. 1894) demonstrated that at moderate pressures helium remains in the liquid state no matter how close its temperature is to absolute zero. All other substances become solids at much higher temperatures. Helium solidifies only at pressures above several tens of atmospheres (see Fig. 14.5). There is no sublimation line for helium, that is, under no circumstances can solid helium be in a state of equilibrium with its vapour.

Liquid helium has one important property. At temperatures above 2.19 K its properties are those of common liquefied gases, and in this state it is termed *helium I*. When helium at the pressure of its saturated vapour is cooled below 2.19 K, an abrupt change in its properties takes place and it, still a liquid, goes over to a new state termed *helium II*. In this state liquid helium constitutes a sort of mixture of the two components: a normal one (helium I) and a superfluid one perfectly devoid of viscosity. These two components can move freely one inside the other without interaction. The superfluid component flows through tiniest capillaries and slits without any friction at all.

The regions of existence of helium I and helium II in the phase diagram (Fig. 14.5) are separated by a dashed line. The fraction of the superfluid component formed in the course of the helium I to helium II transition increases with a decrease in temperature, so that at absolute zero all the helium should go over to the superfluid state. This is, of course, highly hypothetical, since no substance can be cooled to 0 K.

The phenomenon of superfluidity of helium discovered by Kapitza was explained on the basis of quantum mechanics by the prominent Soviet scientist Lev D. Landau (1908-1968).

According to the quantum theory, the energy of molecules at absolute zero is not zero as the classical kinetic theory asserts (see Section 4-3). Even at absolute zero the molecules possess so-called *zero-point energy*, which is the minimum energy they can possess. The forces of interaction between the helium atoms are very small and zero-point energy of helium proves high enough to prevent the formation of a lattice by the helium atoms. Only applying high external pressure can the helium atoms be brought close enough to be able to form a lattice.

The superfluid component in helium II appearing at temperatures close to the absolute zero consists of helium atoms with zero-point energy.

## Thermal Expansion

## 15

### 15-1 Basic Facts About Thermal Expansion

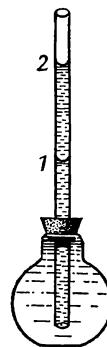
The dependence of the volume of a gas on temperature was established in Chapter 5. Let us consider experiments which prove that all substances and not only gases expand upon heating.

Fill a flask with a liquid (for instance, with water), close the flask with a plug with a glass tube passing through it and note the level of the liquid in the tube (Fig. 15.1). If we heat the liquid in the flask, the level of the liquid in the tube will rise. If the liquid is cooled, the level will drop. An established fact is that liquids expand less than gases.

It is not so easy to observe the thermal expansion of solids. To demonstrate their expansion, a metal ball and a ring through which the ball tightly passes at room temperature can be used. After the ball has been heated in the flame of a burner, it no longer passes through the ring. However, after cooling the ball passes through the ring again. The cause of thermal expansion of solids and liquids was made clear in Section 6-5.

Hence, all substances expand upon heating and contract upon cooling. Let us now see how to measure the magnitude of such expansion.

Fig. 15.1 As the flask is heated, the level of water moves from position 1 to position 2.



## 15-2 Linear Expansion

We recall that crystals are anisotropic and therefore the magnitude of a crystal's expansion, in general, depends on direction. However, most substances are polycrystalline and, by force of this, isotropic. One should keep in mind that every point made below in this chapter applies only to isotropic bodies.

Thus, isotropic solids expand equally in all directions. In practice only unilateral expansion is often of importance. For instance, when pipes are laid for steam pipe lines, only the elongation of the pipes has to be taken into account, since the increase in the cross section is of no practical importance. The variation of a definite dimension of a solid body caused by temperature variations is termed *linear expansion* (*linear contraction*).

Suppose we have a rod whose length at  $0^\circ\text{C}$  is  $l_0$  and at temperature  $t$  is  $l_t$ . This means that the variation of the rod's length upon heating by  $\Delta t = t - 0 = t$  is  $\Delta l = l_t - l_0$ . It can easily be established that the variation of the rod's length  $\Delta l$  is directly proportional to the temperature increment  $\Delta t$  and to its length at  $0^\circ\text{C}$ :

$$\Delta l = \alpha l_0 t \quad (15.1)$$

The quantity  $\alpha$  characterizing the dependence of the linear expansion on the substance and on the external conditions is termed the *coefficient of linear expansion*. It shows the relative variation of the length of a body initially at  $0^\circ\text{C}$  heated or cooled by  $1^\circ\text{C}$ :

$$\alpha = \Delta l / l_0 t \quad (15.1a)$$

(Show that the dimensionality of  $\alpha$  is  $^\circ\text{C}^{-1}$ .)

Let us find the formula which will enable us to calculate the length of a body at different temperatures from its length  $l_0$  at  $0^\circ\text{C}$ . From (15.1) we have

$$l_t - l_0 = \alpha l_0 t$$

or

$$l_t = l_0 (1 + \alpha t) \quad (15.2)$$

To find the length of a body  $l_2$  at a temperature  $t_2$  from its length  $l_1$  at a temperature  $t_1$ , one should actually first find  $l_0$  using formula (15.2) and then apply the same formula to calculate  $l_2$ . However, since  $\alpha$  is a very small number (for instance, for copper  $\alpha = 1.7 \times 10^{-5} \text{ }^\circ\text{C}^{-1}$ ), to find  $l_2$  the following approximate formula may be used:

$$l_2 \approx l_1 [1 + \alpha (t_2 - t_1)] \quad (15.3)$$

From formula (15.3) we obtain an approximate formula for calculating the linear expansion coefficient of a solid:

$$\alpha \approx \frac{l_2 - l_1}{l_1 (t_2 - t_1)} \quad (15.3a)$$

### 15-3 Volume Expansion of Heated Bodies

Let us now consider the relations expressing the dependence of the volume of a body on its temperature.

Let the volumes of a body at  $0^\circ\text{C}$  and at  $t^\circ\text{C}$  be  $V_0$  and  $V_t$ , respectively. In this case the volume variation corresponding to a rise in temperature by  $\Delta t = t - 0 = t$  is  $\Delta V = V_t - V_0$ . Experiments show this variation of the volume of the body to be directly proportional to the temperature increment and to the initial volume  $V_0$ :

$$\Delta V = \beta V_0 t \quad (15.4)$$

The quantity  $\beta$  characterizing the dependence of the volume expansion of a heated body on its substance and on the external conditions is termed the *coefficient of volume expansion*. It shows the relative variation of the volume of a body initially at  $0^\circ\text{C}$  heated or cooled by  $1^\circ\text{C}$ :

$$\beta = \Delta V / V_0 t \quad (15.4a)$$

(Show that the dimensionality of  $\beta$  is  $^\circ\text{C}^{-1}$ .) In calculations  $\beta$  may be assumed to be constant; its values should be taken from tables.

The temperature dependence of the volume of a body may be easily obtained from formula (15.4):

$$V_t - V_0 = \beta V_0 t, \text{ or } V_t = V_0 (1 + \beta t) \quad (15.5)$$

If the volume of a body  $V_1$  at a temperature  $t_1$  is known, its volume  $V_2$  at a temperature  $t_2$  may be found from the following approximate formula:

$$V_2 \approx V_1 [1 + \beta (t_2 - t_1)] \quad (15.6)$$

From (15.6) we obtain an approximate formula for calculating the volume expansion coefficient:

$$\beta \approx \frac{V_2 - V_1}{V_1 (t_2 - t_1)} \quad (15.6a)$$

Note that all the formulae of this section are valid only if the mass of the body,  $m$  remains constant with the change in its temperature. This means that the density of a substance must be temperature-dependent, since its volume changes with temperature.

Indeed, the expression for the density of a substance at  $0^\circ\text{C}$  is  $\rho_0 = m/V_0$  and at a temperature  $t$  it is  $\rho_t = m/V_t$ . Substituting into the latter formula the expression for  $V_t$  (15.5), we obtain

$$\rho_t = \frac{m}{V_0(1+\beta t)} = \frac{m}{V_0} \frac{1}{1+\beta t} = \rho_0 \frac{1}{1+\beta t} \quad (15.7)$$

In calculations one should take into account that the densities of substances shown in tables are for  $0^\circ\text{C}$ . Densities at temperatures other than  $0^\circ\text{C}$  should be calculated with the aid of formula (15.7). Note that the density of a substance decreases upon heating and increases upon cooling.

### 15-4 Thermal Expansion of Solids

Let us demonstrate that for solids there is a simple relation between the coefficients of volume expansion  $\beta$  and of linear expansion  $\alpha$ .

Fig. 15.2 Expansion of heated cube.

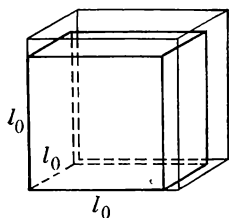


Figure 15.2 shows a solid cube of material with expansion coefficient  $\alpha$  and  $\beta$  whose edge is  $l_0$  at  $0^\circ\text{C}$ . We can write

$$V_0 = l_0^3$$

After the cube had been heated to a temperature  $t$ , its edge will be  $l_t = l_0(1 + \alpha t)$  long and its volume will be  $V_t = l_t^3$ . Hence

$$V_t = l_t^3 = l_0^3(1 + \alpha t)^3$$

On the other hand

$$V_t = V_0(1 + \beta t) = l_0^3(1 + \beta t)$$

and hence

$$l_0^3(1 + \beta t) = l_0^3(1 + \alpha t)^3$$

or

$$(1 + \beta t) = 1 + 3\alpha t + 3\alpha^2 t^2 + \alpha^3 t^3$$

If we take into account that  $\alpha$  is very small, we can neglect terms with  $\alpha^2$  and  $\alpha^3$  and obtain  $\beta t = 3\alpha t$  whence

$$\beta = 3\alpha \quad (15.8)$$

It is now evident that knowing the coefficient of linear expansion,  $\alpha$ , we can use formula (15.8) to compute the numerical values of the coefficient of volume expansion,  $\beta$ . Therefore in practice only the coefficients of linear expansion for solids are included in tables. Because of that it is

reasonable to write formula (15.5) for solids in the form

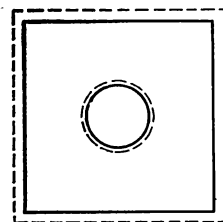
$$V_t = V_0 (1 + 3\alpha t) \quad (15.9)$$

It may easily be inferred that to find the area of a solid's surface one can make use of the formula

$$A_t = A_0 (1 + 2\alpha t) \quad (15.10)$$

where  $A_0$  is the area of this surface at  $0^\circ\text{C}$ . Note that when a homogeneous solid body of arbitrary shape is heated, the distance between any of its two points increases in accordance with formulae (15.2) or (15.3). For instance, a hole in a metal sheet (Fig. 15.3) increases exactly like the circle of identical diameter drawn on an intact sheet. Hence, holes and cavities inside a solid experience the same changes when the solid is heated as if they were completely filled with the solid's material. (How will the gap between the ends of a rod bent in the shape of a ring change when the rod is heated?)

Fig. 15.3 Hole in heated metal sheet widens.



### 15-5 Thermal Expansion of Liquids

In Section 15-1 an experiment involving heating a liquid in a flask was described. We saw that a liquid expands upon heating. Consequently, a liquid expands more than the flask, otherwise its level in the flask would not have risen.

Comparing the coefficients of volume expansion, one sees that liquids expand upon heating by one or sometimes even two orders of magnitude more than solids. For this reason in calculating effects of heating of liquids the expansion of vessels containing them is usually neglected.

In more rigorous calculations one should also take into account the expansion of the heated vessel,  $\Delta V_v$ . Let us call the thermal expansion of a liquid determined from the variation of its level in a vessel *apparent expansion* and denote it by  $\Delta V_{\text{liq. a}}$ . Then the true expansion of the liquid,  $\Delta V_{\text{liq}}$ , should be equal to the sum of the expansion of the inner volume of the vessel containing the liquid and the apparent expansion of the liquid:

$$\Delta V_{\text{liq}} = \Delta V_{\text{liq. a}} + \Delta V_v \quad (15.11)$$

Note that among the liquids there is a remarkable exception to the general rule: water heated from  $0^\circ\text{C}$  to  $4^\circ\text{C}$  contracts and cooled from  $4^\circ\text{C}$  to  $0^\circ\text{C}$  expands. The cause of such anomalous behaviour of water was explained in Section 14-3. Also,  $\beta$  for water changes substantially with temperature.

In the range from 5 °C to 10 °C,  $\beta_{\text{water}} = 0.000\,053$ , and in the range from 60 °C to 80 °C,  $\beta_{\text{water}} = 0.000\,59$ , that is, it changes by a factor of ten.

### 15-6 Thermal Expansion in Nature and Technology

The expansion of bodies upon heating and their contraction upon cooling play an enormous part in nature. Nonuniform heating of air at ground level creates convection currents (wind), leading to changes in weather. Nonuniform heating of water in seas and in oceans creates currents that affect the climate of coastal countries. Especially drastic temperature variations take place in mountain areas. They are the cause of successive expansion and contraction of rock. Since such variations of volume are different for different minerals constituting the rock, variations of temperature create cyclic stresses in the rock that lead to the formation of cracks in them; in the course of time these cracks cause the breaking up of the rock.

The temperature dependence of density, length and of volume of materials is also very important in everyday life and technology. The temperature dependence of air density is used in homes to obtain uniform distribution of heat produced by stoves and heaters, in stoves to create drafts, and in refrigerators for uniform cooling of the chamber.

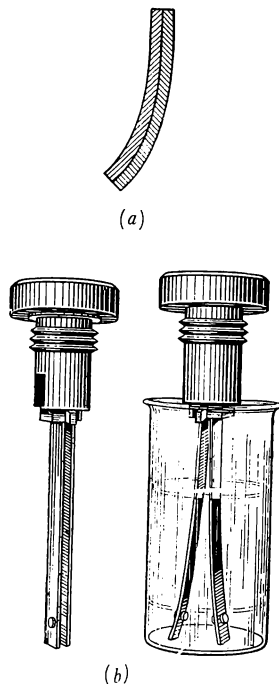
Various automatic systems employ bimetallic strips. Such a strip consists of two strips of different metals bonded together. When the bimetallic strip is heated, one of the strips expands more than the other causing the strip to bend (Fig. 15.4). Such strips are used for automatic on-and-off switching of electric circuits in thermostats, refrigerators, fire alarms, etc.

The temperature dependence of length has to be taken into account when tensioning wires for railways and for electric power lines, in laying steam pipe lines in plants, in building bridges, etc.

In the fabrication of metal-to-glass seals for, say, electric lamps and radio tubes, metals and glass with close values of thermal expansion coefficients are used.

This is by no means a complete list of examples of the part played by thermal expansion in nature and technology.

**Fig. 15.4** (a) Bimetallic strip bends upon heating; (b) contact between bimetallic strips is broken upon heating.



part two

# **Electricity and Magnetism**



# The Fundamentals of the Electron Theory of Atomic Structure. Coulomb's Law

## 16-1 Electrification of Bodies. The Concept of an Electric Charge

From everyday experience we know that many objects after being rubbed begin to attract specks of dust, pieces of paper, hair, etc. Such attraction is due to the *electric charges* produced by the rubbing and accumulated on their surfaces. For this reason we speak of *electrification of bodies*.

Experience has shown that between two electrified bodies there can be either *attraction* or *repulsion*. The explanation is that the electric charges can be of two types. Let us call one of them *positive* and the other *negative*. It was established earlier that *like charges repel* and *unlike charges attract* (Fig. 16.1). The forces of interaction between charges and between electrified bodies (due to the presence of electric charges on them) are termed *electric forces*. It should be kept in mind that electric forces act only on bodies carrying electric charges or on charged particles (for instance, ions).

A body can be electrified not only by rubbing, but also by bringing it into contact with a charged body. Thus, a paper cartridge suspended on a nonconducting silk thread after touching a charged rod is repelled by it (Fig. 16.2). This

means that in the course of contact the charges from the electrified rod partially went over to the cartridge.

Electrify two identical cartridges with charges of unlike sign and let them touch (Fig. 16.3a). After contact the attraction between the cartridges will either cease altogether (Fig. 16.3b) or the cartridges will be repelled (Fig. 16.3c). In the first case *neutralization* of the charges is said to have taken place. Experiment has shown that in such cases the charges are not destroyed but are simply redistributed so that we are no longer able to detect their presence. In such a situation it is natural to assume that a body contains equal numbers of positive and negative charges.

So the neutralization of the cartridges takes place in the course of contact if their charges before contact were equal in magnitude. If they were not equal, the smaller charge neutralizes the part of the opposite charge equal to it in magnitude, and the remaining part of the greater charge is shared between both cartridges so that they begin to repel each other.

The excess of electric charges of like sign in a body is termed the *quantity of electricity* of the body. The numerical value of the quantity of electricity a body contains can be found from the force with which it interacts with other electrified bodies. The total charge of any body is the algebraic sum of all electric charges contained in the body.

Note finally that no matter what method of electrification is employed the electric charges are neither produced nor destroyed, but only redistributed among the bodies taking part in the phenomenon being considered. This rule is known as the *law of charge conservation*.

## 16-2 The Complex Nature of the Atomic Structure

The phenomenon of the electrification of bodies shows clearly that electric charges are part of matter. However, for a long time there was no answer to the problem of the part they play in matter and in its structure. In the second half of the nineteenth century many phenomena had been discovered pointing to the complex structure of the atom. The analysis of those phenomena has led to the conclusion that electric charges must be part of the atom.

Since matter in its natural state is, generally speaking, electrically neutral, it could be assumed that atoms contain positive and negative electric charges of equal magnitude. The problem of the nature of these charges and the ways they are held together in the atom was an urgent one. By the

Fig. 16.1 Interaction of like and of unlike charges.

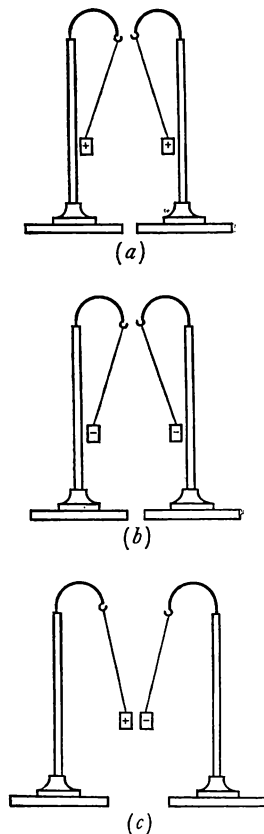
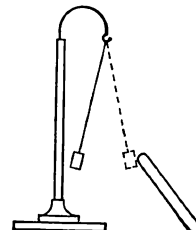
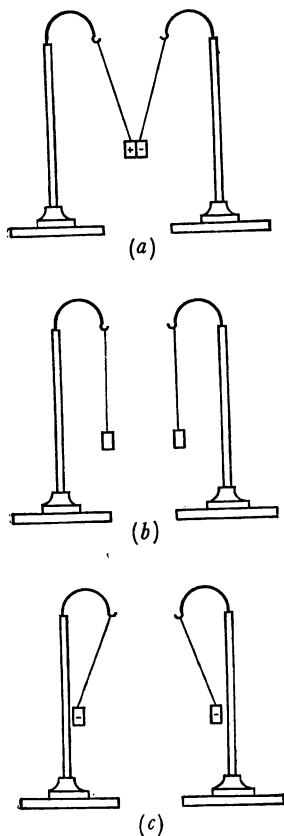


Fig. 16.2 Electrification by contact.



**Fig. 16.3** (a) Cylinders carrying unlike charges exchange them upon contact; (b) when cylinders make contact, charges are neutralized; (c) cylinders repel each other after being in contact because negative charge was greater than positive.



beginning of the twentieth century these questions became a major scientific problem, the solution of which critically affected future progress in physics, chemistry and allied sciences.

The complex nature of the atomic structure could be inferred primarily from various chemical reactions. The formation of molecules from atoms and the rearrangement of molecules in the course of chemical reactions were proof of the existence of intermolecular attraction, which could be explained by the presence of unlike electric charges in the atoms.

The laws of electrolysis discovered by Faraday (see Chapter 24) demonstrated that substances must contain indivisible (elementary) electric charges, apparently identical for all atoms.

The all-important works of Mendeleev brought to light periodicity in the properties of chemical elements, which could be explained by the recurrence of combinations in the arrangement of charges inside atoms.

In experiments with rarefied gases (see Section 23-4), in the photoeffect (see Section 38-6) and in some other cases negatively charged particles were discovered for which the Irish physicist George J. Stoney (1826-1911) proposed the term *electrons*.

The French scientist physicist A. H. Becquerel (1852-1908) in 1896 established that uranium ore emits radiation of complex composition, while the British scientist physicist Lord Ernest Rutherford (1871-1937) later proved that it consisted of three types of radiation, called alpha-rays ( $\alpha$ -rays), beta-rays ( $\beta$ -rays) and gamma-rays ( $\gamma$ -rays).

The *alpha-rays* turned out to be positively charged and are now known to consist of the nuclei of helium atoms ( $\alpha$ -particles) flying at speeds of the order of  $10^7$  m/s. *Beta-rays* are electrons flying at speeds of the order of  $10^8$  m/s. *Gamma-rays* turned out to be very short electromagnetic waves.

All these discoveries served as a basis for various models of atoms made up of electric charges, but the true structure of atoms was discovered only in 1911 by Rutherford in the course of his studies of the scattering of  $\alpha$ -particles by substances.

### 16-3 Rutherford's Experiment and the Nuclear Idea

Rutherford studied atomic structure in experiments in which he placed metal foil in the path of  $\alpha$ -particles flying in a definite direction. He then recorded the scattering of

$\alpha$ -particles by the foil. Rutherford established that the great majority of  $\alpha$ -particles continued on their original course or were deflected by a very small angle. Only a small proportion of the  $\alpha$ -particles changed their course appreciably, and some particles bounded back. This proved that almost all of the atom's mass was concentrated in its centre, that is, that there is a very small, but massive, positively charged particle in the centre of the atom that repels  $\alpha$ -particles flying close to it and can even reject an  $\alpha$ -particle flying head on at it. The positive charged particle in the centre of the atom was termed the *atomic nucleus*.

Since an atom in its normal state is electrically neutral, it should also contain negative charges to compensate for the positive charge of the nucleus. Since the electrons are attracted to the nucleus, they should fall on it. Accordingly, Rutherford assumed that the electrons in an atom move about the nucleus in circular or elliptical orbits, so that the force of their attraction to the nucleus is counterbalanced by centrifugal force. The model of atomic structure proposed by Rutherford basically looks like the structure of the solar system, and because of that it has been termed the *planetary*, or *nuclear*, model.

Thus, according to Rutherford every atom consists of a small, but massive, positively charged nucleus with negatively charged electrons orbiting it. Subsequent research proved that the nuclear model proposed by Rutherford was substantially true.

It has been established that the smallest nucleus was that of the hydrogen atom. The term for it is *proton*. The radius of the proton is approximately  $10^{-15}$  m and its mass is  $1.672 \times 10^{-27}$  kg. Only one electron orbits this nucleus. The electron's mass is  $9.11 \times 10^{-31}$  kg, which is approximately 1836 times less than the proton's mass. Hence, almost all of the hydrogen atom mass is concentrated in its nucleus, and the same is true of the atoms of all the other elements.

#### 16-4 The Atomic Structure of Chemical Elements

It has been established that the positive charges of the atomic nuclei of all chemical elements contain integral numbers of the proton charge. Research proved that there are protons in all atomic nuclei. It was also established that the chemical nature of the atom is uniquely determined by the magnitude of the positive charge of its nucleus, that is, by the number of protons contained in it. This number was termed the *atomic number*,  $Z$ .

All atoms with an identical atomic number for their nucleus are atoms of the same chemical element, even though their masses may prove different. In the Mendeleev Periodic Table all the chemical elements are arranged in order of atomic numbers. Accordingly, all atoms with identical atomic numbers but with different masses occupy the same position in the table. Such atoms are termed *isotopes* (from the Greek *topos* for place).

In every neutral atom there should be  $Z$  electrons orbiting the nucleus, that is, their number should be equal to the number of protons in the nucleus.

The electrons in the atoms are arranged in layers called *shells*. The electrons of the outermost shell, that is, the most distant from the nucleus, are termed *valence electrons*. The maximum number of such electrons is eight. The valence electrons are comparatively weakly bonded to the nucleus and when acted upon by external forces can leave their atoms and go over from one atom to another.

The atoms of metals with one to three electrons in their outermost shell give up their valence electrons more readily than others (see Section 13-3). This is the reason why all metals are good conductors. The atoms of metalloids, whose atoms have from five to seven electrons in their outermost shell, exhibit a tendency to annex additional electrons to bring their number up to eight. For this reason all metalloids are poor conductors.

Atoms and molecules with a surplus or shortage of electrons compared to their normal state are termed *ions*. The term for the process of acquiring surplus electrons or tearing electrons away from neutral atoms is *ionization*. Atoms and molecules which acquire surplus electrons become negatively charged and are termed *negative ions*, while atoms and molecules which lose electrons and thereby become positively charged are called *positive ions*.

An electron in an atom or molecule is said to be *bound*. An electron torn away from some atom and remaining as an individual particle, outside any other molecule or atom is termed *free*. In all cases the electrification of a body involves the acquisition or loss of electrons or ions by the body.

Note that the arrangement of electrons inside atoms or molecules plays an important part in determining their physical and chemical properties. Studies of the arrangement of electrons inside the atoms of various chemical elements lead to the conclusion that the maximum number of electrons in a shell numbered  $n$  is  $2n^2$ . Therefore the first (from the centre) filled shell must contain 2 electrons, the

second 8, the third 18, etc. These shells are often denoted by the capital letters  $K$ ,  $L$ ,  $M$ , etc (Fig. 16.4).

The building of molecules from atoms brought close together is mainly determined by the behaviour of the valence

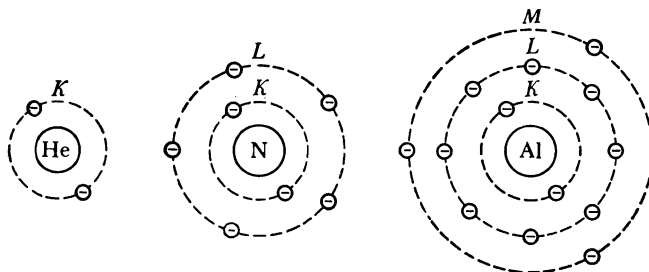


Fig. 16.4 Schematic representation of helium (He), nitrogen (N) and aluminium (Al) atoms; electrons are arranged in shells around nuclei.

electrons of those atoms. As an example, let us see how the ionic bond is established. Suppose we have neutral potassium and chlorine atoms (Fig. 16.5*a*). When a pair of such atoms comes together, the only electron in the potassium

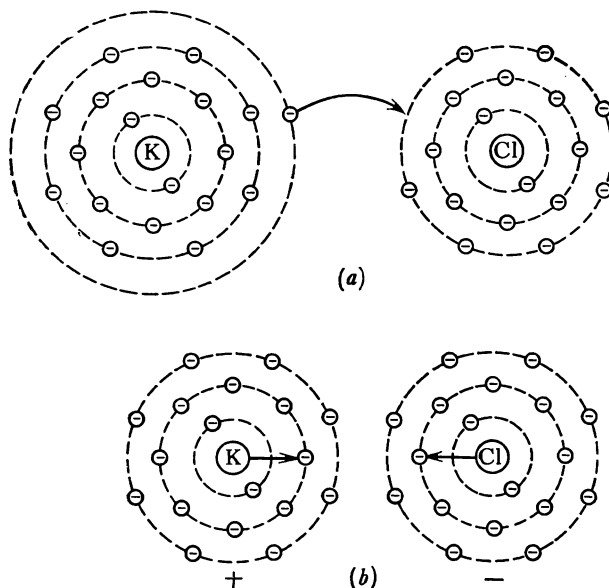


Fig. 16.5 Schematic representation of process of formation of potassium chloride molecule: (a) electron moves over from potassium atom to chlorine atom; (b) ions of potassium and chlorine are attracted and form molecule of KCl.

atom's outermost shell goes over to the chlorine atom, the atoms of potassium and chlorine thus becoming positively and negatively charged ions respectively with 8 electrons each in their outermost shells, this number corresponding to the most stable arrangement of electrons in an atom.

The oppositely charged ions of potassium and chlorine are attracted and form the KCl molecule. In a potassium chloride crystal the alternating ions of potassium and chlorine form an ionic crystal lattice. Since every potassium ion interacts at the same time with several chlorine ions, there are strictly speaking no molecules. When such a lattice breaks up not molecules but ions split off it. It is because of this that aqueous solutions of ionic crystals are good conductors of electricity (see Section 21-4).

Note in addition that only electrons act as mobile charge carriers in solid metals. Consequently only electrons are transported in the process of the electrification of metal bodies by contact. The theory which makes use of electron motion to explain electrical phenomena is termed *electron theory*.

### 16-5 Electrification by Contact

Since the number and arrangement of electrons are different in different atoms, the forces retaining valence electrons in them depend on their nature. Therefore, when different atoms are in contact, the electrons may go over from the atom which attracts them less to the atom which attracts them more (see Section 16-4).

Hence, when two bodies are in contact, some electrons will go over from one body to another, that is, the bodies will become electrified. Their charges will be equal in magnitude and opposite in sign. To separate such bodies work will have to be performed against the forces of electrical attraction. The charged bodies gain energy at the expense of this work.

The magnitude of the charges of such bodies obviously depends on the properties of the bodies and on the area of their contact. Naturally, when bodies of identical material are brought in contact there will be no electrification.

If different bodies are simply pressed together, the points of contact are few in number. Rubbing increases the contact area substantially and thereby greatly increases electrification.

### 16-6 Interaction Between Electric Charges. Coulomb's Law

The electric charge of a proton is usually designated  $e_+$  and that of an electron  $e_-$ . These charges are termed *elementary* because no one has been able to detect smaller charges-

The point will be discussed in more detail in Section 17-18.

Since the protons and electrons in all atoms have charges of identical magnitude,  $e$ , the total electric charge of any body  $q$  may be expressed in the following form:

$$q = (N_+ - N_-)e \quad (16.1)$$

where  $N_+$  is the number of particles with the charge  $e_+$ , and  $N_-$  the number of particles with the charge  $e_-$ . If  $N_+ = N_-$ , the body is neutral and  $q = 0$ , and if  $N_+ \neq N_-$ , the body has a net charge.

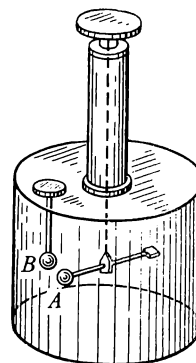
The force of interaction between electric charges can be measured with the aid of a *torsion balance* (Fig. 16.6). This instrument has two small metal balls on insulating rods. One of the rods with the ball  $B$  is stationary, while the other carrying the ball  $A$  and a counterweight on its other end is suspended on a thin elastic thread. The upper end of the thread is attached to a knob which can be rotated to increase or decrease the distance between balls  $A$  and  $B$ . The balls are charged and the force of interaction of their charges is determined from the angle through which the thread twists. Note that the force of gravity acting between the balls can be ignored because it is many times less than the electric force.

After the force of interaction between two given charges has been determined, the distance between the balls is changed and the force measured again. It turns out that when the distance is increased two, three or four times the force of interaction decreases four, nine, or sixteen times respectively, that is, changes in inverse proportion to the square of the distance. The magnitude of the charges can be altered in the following way. If we touch one of the charged balls, say  $B$ , with an identical uncharged ball, the charge  $q$  will be equally divided between the two, leaving the charge  $q/2$  on the ball  $B$ . Such division of the charge can be continued. If we continue to decrease the charge of one of the balls two, three, or four times with the distance between the balls  $A$  and  $B$  remaining constant, the force of interaction between the charges will decrease the same number of times. This means that the force of interaction is directly proportional to the magnitude of the charge of each of the balls.

All these conclusions are valid as long as the dimensions of the bodies carrying the charges remain small in comparison with the distance between them. Such charges are termed *point charges*.

The experiments we have just described were first carried out by the French physicist Charles A. Coulomb (1736-1806). He established a law which now carries his name, Coulomb's

Fig. 16.6 Torsion balance.





law: the force of interaction between two point charges is directly proportional to the product of their magnitudes, is inversely proportional to the square of the distance between them and is directed along a straight line connecting the charges:

$$F = K \frac{q_1 q_2}{r^2} \quad (16.2)$$

Here  $r$  is the distance between the charges, and  $K$  is a proportionality factor.

### 16-7 The Permittivity of a Medium

Since electric charges are part of all molecules, the medium surrounding charged bodies could be expected to affect the electric force of the interacting bodies. Experiment proves this conjecture to be true.

The force of interaction between two charges is at its maximum in a vacuum. A medium always reduces the force. This means that the factor  $K$  in (16.2) depends both on the choice of units and on the properties of the medium. Therefore  $K$  is conveniently expressed in the form  $K = k/\epsilon_m$ , where  $k$  depends on the choice of units, and  $\epsilon_m$  characterizes the medium. The quantity  $\epsilon_m$  is termed the *permittivity of the medium*.

Hence formula (16.2) assumes the form:

$$F = k \frac{q_1 q_2}{\epsilon_m r^2} \quad (16.3)$$

For the force of interaction between the same charges in a vacuum,  $F_0$  (16.3) should be written in the form

$$F_0 = k \frac{q_1 q_2}{\epsilon_0 r^2} \quad (16.3a)$$

where  $\epsilon_0$  is called *permittivity of free space*. Dividing (16.3a) by (16.3) we obtain

$$\frac{F_0}{F} = \frac{\epsilon_m}{\epsilon_0} = \epsilon$$

The term for  $\epsilon = \epsilon_m/\epsilon_0$  is *relative permittivity*, or *dielectric constant*. Its value is a dimensionless number always greater than unity since  $\epsilon_m > \epsilon_0$ . The relative permittivity of a medium is the ratio of the force of interaction of specific charges in a vacuum to that in the medium. The numerical value of  $\epsilon$  is determined from experiments and is then obtained from tables.

The permittivity of a medium,  $\epsilon_m$ , is expressed by the formula

$$\epsilon_m = \epsilon \epsilon_0 \quad (16.4)$$

Obviously, the numerical value of  $\epsilon_0$  should be previously measured in experiment or otherwise agreed on.

### 16-8 SI Units in Electricity

The International System of Units is designed so as to make the factor  $k$  in Coulomb's law (16.3) equal to  $1/4\pi$ , where  $\pi = 3.141 \dots$ . Therefore in the SI system Coulomb's law takes the form

$$F = \frac{q_1 q_2}{4\pi \epsilon_m r^2} \quad (16.5)$$

$$F = \frac{q_1 q_2}{4\pi \epsilon_0 \epsilon r^2} \quad (16.5a)$$

Since one of the base units (see Section 1-9) in the SI system is the unit of electric current, the ampere, the unit of charge in this system is a derived one, being a result of the formula

$$q = It \quad (16.6)$$

For a unit charge

$$q = 1 \text{ A} \cdot 1 \text{ s} = 1 \text{ A} \cdot \text{s} = 1 \text{ C (coulomb)}$$

Thus, the unit of charge in the SI system is the *coulomb*. Coulomb is the term for the electric charge transported through a conductor's cross section by a current of 1 A in the time of 1 s. The coulomb is a very large unit of charge, being equal to the charge of  $625 \times 10^{16}$  protons (electrons). Experiments have shown (see Section 17-18) that  $e_+ = 1.60 \times 10^{-19} \text{ C}$  and  $e_- = -1.60 \times 10^{-19} \text{ C}$ . (Ponder on the question of how the number of protons cited above was determined.) The unit for  $\epsilon_m$  may be expressed in the SI system in the form

$$\epsilon_m = \frac{q_1 q_2}{4\pi r^2 F}, \quad \epsilon_m = \frac{1 \text{ C} \times 1 \text{ C}}{4\pi \times 1 \text{ m}^2 (1/4\pi) \text{ N}} = 1 \frac{\text{C}^2}{\text{N} \cdot \text{m}^2}$$

In the SI system the unit of permittivity is the permittivity of a medium in which two charges of one coulomb each, a distance of 1 m apart, interact with a force equal to  $1/4\pi$ -th of a newton.\*

\* This unit in the SI system bears the name *farad per metre* (F/m), see Section 17-13.

The following value of  $\epsilon_0$  was obtained from experiments:

$$\epsilon_0 = \frac{1}{36\pi \times 10^9} \frac{\text{C}^2}{\text{N} \cdot \text{m}^2} = 8.85 \times 10^{-12} \frac{\text{C}^2}{\text{N} \cdot \text{m}^2} \quad (16.7)$$

This value of  $\epsilon_0$  is called the *electric constant*.

### 16-9 Gaussian Units in Electrostatics

The Gaussian system of units may also be used in electricity. It is the CGS system applied to electricity. In this system the factor  $k$  in formula (16.3) is assumed to be equal to unity and the permittivity of free space is also regarded as a dimensionless unity ( $\epsilon_0 = 1$ ). Therefore, in the Gaussian system Coulomb's law is written in the form

$$F = \frac{q_1 q_2}{\epsilon r^2} \quad (16.8)$$

where  $\epsilon$  is the relative permittivity. This formula can be used to derive the unit of charge  $q$ . Assuming  $q_1 = q_2 = q$ , we obtain

$$F = \frac{q^2}{\epsilon r^2}, \quad \text{whence} \quad q = r \sqrt{\epsilon F}$$

Accordingly

$$\begin{aligned} q &= 1 \text{ cm} \sqrt{1 \times 1 \text{ dyn}} = 1 \text{ cm} \left( \frac{\text{g} \cdot \text{cm}}{\text{s}^2} \right)^{1/2} \\ &= 1 \frac{\text{cm} \cdot \text{g}^{1/2} \text{ cm}^{1/2}}{\text{s}} = 1 \frac{\text{g}^{1/2} \text{ cm}^{3/2}}{\text{s}} = 1 \text{ statC (statcoulomb)} \end{aligned}$$

The unit charge in the Gaussian system is a charge acting on an equal charge a distance of 1 cm away with a force of 1 dyne in a vacuum. Note that

$$1 \text{ C} = 3 \cdot 10^9 \text{ statC}$$

Fig. 16.7 Electroscope.



The numerical values of electron and proton charges in the Gaussian system are

$$e_+ = 4.8 \times 10^{-10} \text{ statC} \quad \text{and} \quad e_- = -4.8 \times 10^{-10} \text{ statC}$$

### 16-10 The Electroscope

The instrument used for the detection of an electric charge is called an *electroscope*. The electroscope of the simplest kind consists of a metal rod on one end of which there is a ball and on the other, two strips of metal foil (Fig. 16.7). The rod passes through a dielectric plug into a plastic or

glass vessel, which protects the strips from the effects of air circulation. If a charged body touches the ball, the strips move apart, since they carry like charges.

In some experiments it is desirable to have an electroscope carrying a predetermined charge. In such cases it is charged by touching it with a glass rod previously rubbed against leather (the glass acquires a positive charge) or with an ebonite rod rubbed against fur (the ebonite acquires a negative charge). Such a charged electroscope can be used to determine the charge of any electrified body. (How can this be done?)

## The Electric Field

## 17

### 17-1 Electric Field as a Special Form of Matter

Signals from distant events always reach us through some medium. For instance, telephone communication is realized with the aid of electric wires, speech is transmitted by sound waves propagating in air (sound cannot propagate in a vacuum), and so on. Since the appearance of a signal is a material phenomenon, its propagation, involving the transfer of energy from point to point in space, can take place only in a material medium.

The most important proof that a material medium is involved in the transport of a signal is the finite velocity of its propagation from the source to the observer, this velocity being dependent on the properties of the medium. For instance, sound propagates in air at a speed of about 330 m/s.

If the circumstances existed in which signals travelled at infinite velocity, this would mean that the bodies could interact at a distance, even without any matter between them. In physics such an interaction is termed *action at a distance*. The interaction of bodies via a medium is explained in terms of the *field concept*. Here the bodies act directly on the medium and the medium, in turn, acts on the second interacting body.

It takes time to transmit the action of one body to another through a medium, because all processes in a material medium are transmitted from point to point at a definite velocity. The mathematical theory of fields in physics was developed by the eminent British physicist James Clark Maxwell (1831-1879). Since there are no signals in nature

which are transmitted instantly, we shall in the following discussion keep to the field theory.

In some cases the propagation of signals takes place in a substance, for instance, when sound propagates in air. In other cases no substance is involved in the transmission of the signal. For instance, light from the Sun reaches the Earth through a vacuum. Therefore matter does not exist only in the form of substance.

In cases when the bodies can interact in a vacuum the material medium transmitting this interaction is also called a *field*. Hence, matter exists in the form of substances and fields. The fields may be of different types, depending on the type of forces acting between the bodies. The field which in accordance with the law of universal gravitation transmits gravitational interaction between bodies is termed the *gravitational field*. The field which transmits the interaction between static charges in accordance with Coulomb's law is termed *electrostatic*, or *electric*, *field*.

Experiments have shown that electric signals propagate through a vacuum at a very high but nevertheless finite speed equal to about 300 000 km/s (see Section 30-6). This proves that an electric field is no less a reality than a substance. Research into the properties of the field enabled the transmission of energy through space to be realized and the energy to be used for the benefit of mankind. Radiocommunications, television and lasers serve as examples. However, many properties of fields have been studied incompletely or are still unknown. Research into the properties of fields and the interaction between field and substance is one of the major scientific problems of modern physics.

Any electric charge  $q$  sets up an electric field in space through which it interacts with other charges. An electric field acts only on electric charges. Therefore there is only one method of detecting such a field: to introduce a *test charge*  $q_{\text{test}}$  into space at the point of interest to us. If there is an electric field at this point there will be an electric force acting on  $q_{\text{test}}$ . In the absence of a field this force will be equal to zero.

When a test charge is used to examine a field, it is assumed that it does not distort the field by its presence. This means that the test charge should be very small in magnitude as compared with the charges that set up the field. It has been generally agreed that positive charge be used as test charges.

It follows from Coulomb's law that the magnitude of the force of interaction between electric charges decreases with the distance between them but never vanishes completely.

This means that, theoretically, the field of an electric charge stretches to infinity. However, in practice we take a field to be present only where a perceptible force acts on the test charge.

Note, in addition, that the field of a charge accompanies it in its motion. When the charge moves so far away that the electric force no longer acts on the test charge, we speak of the field as having vanished, although in fact it has simply moved to other points in space.

## 17-2 The Electric Field Strength

Suppose there is an electric point charge  $q$  at some point of space (Fig. 17.1). It follows that there will be an electric force at every point of the surrounding space acting on a test charge. For this reason the field around the charge is sometimes termed a *force field*.

The forces acting on a test charge at different points of the electric field are of different magnitude and direction (see Fig. 17.1). Therefore it is expedient to introduce a special force characteristic for any point of a field set up by an electric charge. It is evident from Coulomb's law that a force  $\mathbf{F}$  cannot serve as such a characteristic because at the same point of the field it is proportional to the magnitude of the test charge  $q_{\text{test}}$ :

$$\mathbf{F} = E q_{\text{test}} \quad (17.1)$$

The proportionality factor in (17.1),  $E$ , remains constant at any point of the field and can serve as a force characteristic of the field. The force characteristic of a point of an electric field  $\mathbf{E}$  is termed the *electric field strength*. Its measure is the force with which the field acts on a positive unit charge placed at this point.

Note that the field strength is a vector coinciding in direction with the force  $\mathbf{F}$  acting on a positive charge  $q_{\text{test}}$  at the specified point of the field, its magnitude being

$$E = F/q_{\text{test}} \quad (17.2)$$

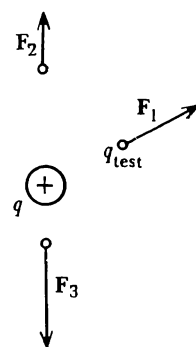
The unit for measuring  $E$  can be deduced from (17.2):

$$E = 1 \text{ N/1 C} = 1 \text{ N/C} = 1 \text{ kg} \cdot \text{m}/(\text{A} \cdot \text{s}^3)$$

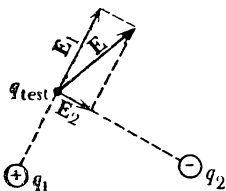
The unit of electric field strength in the SI system can be accepted as the intensity at such a point of an electric field in which a force of 1 N acts on a charge of 1 C.\*

\* This unit of the SI system is termed *volt per metre* (V/m), see Section 17-7.

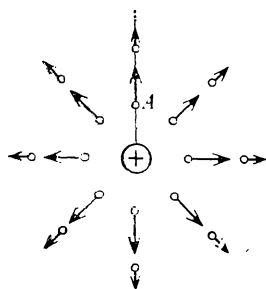
Fig. 17.1 The forces acting on test charge are different at different points in electric field.



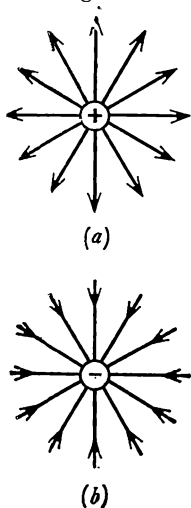
**Fig. 17.2** Strength of field set up by several charges is found by vector summation rule.



**Fig. 17.3** Field strength vectors of positive point charge.



**Fig. 17.4** Representation of field of charge using lines of strength: (a) field of positive charge; (b) field of negative charge.



The field strength is determined by the magnitude and the geometry of the charges that set up the field and can be theoretically calculated. Let us demonstrate how this is done for the case of a separate point charge  $q$ . The force of interaction between the point charges  $q$  and  $q_{\text{test}}$  is

$$F = \frac{qq_{\text{test}}}{4\pi\epsilon_0 r^2}$$

Since  $E = F/q_{\text{test}}$  it follows that

$$E = \frac{q}{4\pi\epsilon_0 r^2} \quad (17.3)$$

Formula (17.3) enables us to determine only the magnitude of vector  $\mathbf{E}$ ; it is directed along the straight line connecting the chosen point of the field with the point charge  $q$ , this direction coinciding with that of the force acting on a positive test charge. If the field is set up by the combined action of several charges, then  $\mathbf{E}$  at some point of the field is determined as the vector sum of the intensities applied at this point by all the charges individually (Fig. 17.2).

Note that in the Gaussian system formula (17.3) takes the form

$$E = \frac{q}{\epsilon r^2} \quad (17.3a)$$

We derive a unit for measuring  $E$  in this system:

$$\begin{aligned} E &= \frac{F}{q}, & E &= \frac{1 \text{ dyn}}{1 \text{ statC}} = \frac{1 \text{ g} \cdot \text{cm/s}^2}{1 \text{ g}^{1/2} \cdot \text{cm}^{3/2}/\text{s}} \\ & & &= 1 \frac{\text{g}^{1/2}}{\text{cm}^{1/2} \cdot \text{s}} = 1 \text{ statV/cm} \end{aligned}$$

(see Section 17-6). One calculates that

$$1 \text{ statV/cm} = 30\,000 \text{ N/C}, \text{ or } 1 \text{ statV/cm} = 3 \times 10^4 \text{ V/m}$$

### 17-3 Electric Field and Lines of Force

One can use vectors (Fig. 17.3) to depict an electric field in space. Such a method of representation takes a lot of time and is not always convenient. Faraday proposed the representation of a field by lines of force for which we will now use the term *lines of strength*. One such line has been drawn through point  $A$  so that vector  $\mathbf{E}$  coincides with its direction. Several straight lines drawn from the charge  $q$  give us a representation of the field in the form of lines of strength (Fig. 17.4).

To differentiate between the representations of fields of positive and negative charges it has been generally agreed to regard the lines of strength as coinciding in direction with

the electric field strength vector  $\mathbf{E}$ . The direction of the lines of strength of fields of positive and negative charges will then be opposite (Fig. 17.4a and b).

It is more difficult to draw the lines of strength when the field is produced by several charges, for instance two charges. In most cases it is impossible to draw a line so that the field strength vectors represented at each of its points lie completely on it. It is, however, possible to draw a curve so that at each point the field strength vectors are tangent to it. Figure 17.5 shows one such line drawn through point  $M$ . (Why only one such line can be drawn through any point of the field?)

Line of strength is the term for a line whose tangents at every point coincide with the field strength vectors. In Figs. 17.6 and 17.7 these lines are used to depict fields of unlike and like charges of equal magnitude.

When making a graphical representation of a field, one should keep in mind that the electric field lines of strength (a) never intersect, (b) start at the positive charge (or infinity) and terminate at the negative charge (or infinity), that is, they are open lines and (c) are not disrupted anywhere between the charges.

The picture of a field arrived at with the aid of lines of strength will be more informative if it is agreed to draw the lines denser in places of greater field strength. Hence, the density of the lines of strength should be made proportional to  $E$ . In calculations it is assumed that the number of lines passing through a unit area perpendicular to them should be numerically equal to  $E$  in the place where that area is chosen. (Consider where the maximum strength is in Figs. 17.6 and 17.7 and what the direction of the field strength vector in three arbitrary points is.)

#### 17-4 The Homogeneous Electric Field

Take two identical metal plates and arrange them parallel to one another as shown in Fig. 17.8. If a charge  $+q$  is imparted to one plate and a charge  $-q$  to the other, an electric field will be established between them. If the distance between the plates is small compared to their linear dimensions, the lines of strength of this field will be parallel.

It can be proved rigorously that for plates of infinite dimensions the density of the lines of strength and their direction will be the same throughout the field between the plates and that consequently

$$E = \text{constant} \quad (17.4)$$

Fig. 17.5 Arrangement of field strength vectors of two equal unlike charges.

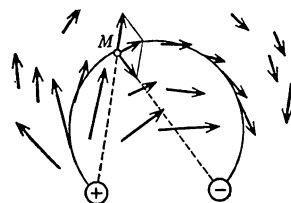


Fig. 17.6 Representation of field of two equal unlike charges.

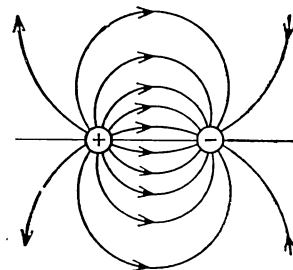


Fig. 17.7 Representation of field of two equal like charges.

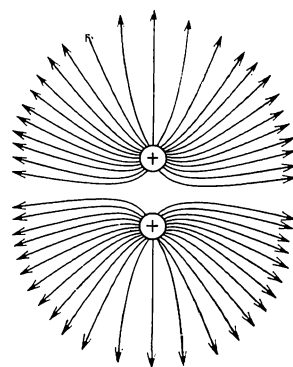
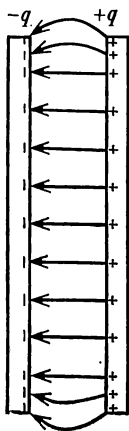




Fig. 17.8 Electric field between plates.



The electric field whose strength vectors are identical at every point is termed *homogeneous*, or *uniform*. Note that only the field in the central part of the system can be regarded as being uniform, since close to the edges vector  $\mathbf{E}$  varies from point to point.

In the case being considered mutual attraction concentrates the entire charges on the internal surfaces of the plates. The quantity  $\sigma$  characterizing the distribution of the charge over the surface is termed *surface charge density*. The measure for the surface density of a charge uniformly distributed over a surface is the amount of electricity per unit area of this surface:

$$\sigma = q/A \quad (17.5)$$

We deduce a unit for measuring

$$\sigma = 1 \text{ C}/1 \text{ m}^2 = 1 \text{ C}/\text{m}^2 = 1 \text{ A} \cdot \text{s}/\text{m}^2$$

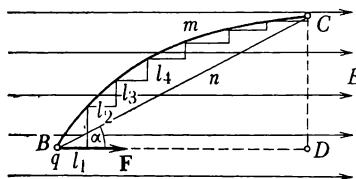
The unit of  $\sigma$  in the SI system is the surface density of a charge of 1 C uniformly distributed over an area of 1 m<sup>2</sup>. Note that this unit is very large (see Section 16-8).

Note in addition that because of the repulsion of like charges the charge density at the edges of metal plates is somewhat greater than at their centres. However, if the distance between the plates is small compared to their linear dimensions, the charge density on the plates away from the edges can be considered constant.

### 17-5 Work Done by an Electric Field in Moving a Charge

Let us see how to calculate the work done by electric forces in displacing a charge  $q$  in a homogeneous electric field ( $\mathbf{E} = \text{constant}$ ). Suppose the charge  $q$  is at point  $B$  in a homogeneous electric field (Fig. 17.9).

Fig. 17.9 Work of forces of electric field in moving charge  $q$  from point  $B$  to point  $C$  is independent of path.



It has been established in mechanics that work is equal to the product of the force, path and cosine of the angle between them. Therefore the work performed by electric forces in displacing a charge  $q$  to point  $C$  along the straight

line  $BnC$  will be expressed as follows:

$$W_{BnC} = F \times BC \times \cos \alpha = qE \times BC \times \cos \alpha$$

Since  $BC \times \cos \alpha = BD$  (see Fig. 17.9), we have

$$W_{BnC} = qE \times BD$$

The work of the field's forces in displacing the charge  $q$  to point  $C$  along the path  $BCD$  is equal to the total work performed along the sections  $BD$  and  $DC$ :

$$W_{BDC} = W_{BD} + W_{DC} = qE \times BD + qE \times DC \times \cos 90^\circ$$

Since  $\cos 90^\circ = 0$ , the work of the field's forces along  $DC$  is zero. Therefore

$$W_{BDC} = qE \times BD$$

This means that when a charge is displaced along a line of strength and then in the direction perpendicular to it the forces of the field perform work only along the line.

Let us now assess the work done along the curve  $BmC$ . Divide the curve into segments so small that every one of them can be regarded as a straight line (see Fig. 17.9). It has been proved above that the work done along each segment is equal to the work done along the corresponding segment of the line of strength  $l_i$ . Then the total work along the path  $BmC$  will be equal to the sum of the works done along  $l_1, l_2, \dots$ . Hence

$$W_{BmC} = qE \times (l_1 + l_2 + \dots + l_h)$$

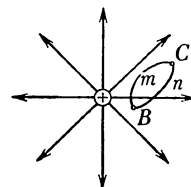
Since the sum is equal to  $BD$ , we have

$$W_{BmC} = qE \times BD$$

Thus we have proved that in a homogeneous electric field the work of electric forces is independent of the path. For instance, when the charge  $q$  is moved from point  $B$  to point  $C$  the work will in all cases be equal to  $qE \times BD$ . It can be proved that this conclusion remains valid for the case of a nonhomogeneous field. This means that, if the distribution of electric charges setting up the field remains constant with time, the forces of the field will be *conservative* (see Section 8-2).

Since the work of the forces of the field along the curves  $BnC$  and  $BmC$  is equal (Fig. 17.10), the work along a closed path is zero. Indeed, if the work is positive along the segment  $BmC$ , it will be negative along the segment  $CnB$ . Hence the work of the forces of an electric field along a closed circuit is always zero.

Fig. 17.10 Work of electric field along closed path is zero.



When only conservative forces are active, work is the sole measure of energy variations. The field of a conservative force, that is, a field in which work is independent of path, is termed a *potential field*. Gravitational and electric fields are potential.

Since the forces of an electric field are conservative, their work in displacing the charge from point  $B$  to point  $C$  (see Fig. 17.10) can serve as a measure of the variation of the potential energy of a charge in an electric field. Denoting the potential energy of the charge at point  $B$  by  $P_B$  and at point  $C$  by  $P_C$ , we write

$$W_{BC} = P_B - P_C \quad (17.6)$$

In the more general case of a charge moving in an electric field from point 1 where its potential energy was  $P_1$  to point 2 where its potential energy is  $P_2$ , the work of the field's forces is

$$W_{12} = P_1 - P_2 = -(P_2 - P_1) = -\Delta P_{21} \quad (17.6a)$$

where  $\Delta P_{21} = P_2 - P_1$  is the increment in the potential energy of a charge displaced from point 1 to point 2.

It follows from (17.6a) that the signs of  $W_{12}$  and  $\Delta P_{21}$  are always opposite. Indeed, if the charge  $q$  is displaced by the forces of the field (i.e., the work of the field,  $W_{12}$ , is positive), its potential energy is reduced (i.e.,  $P_2 < P_1$  and the increment of potential energy is negative). If, on the other hand, the charge is displaced against the forces of the field ( $W_{12} < 0$ ) its potential energy is increased ( $\Delta P_{21} > 0$ ).

It follows from formula (17.6) that by measuring work one can determine only the *difference* between the potential energies of the charge  $q$  at points  $B$  and  $C$  of the field: there are no methods for an unambiguous assessment of the magnitude of its potential energy at a definite point of the field. To eliminate this uncertainty we can agree to accept as zero the potential energy of the charge at any arbitrarily chosen point in the field. In that case the definition of the potential energy at all other points will be unique. It is agreed that the potential energy of a charge at an infinite distance from the charged body that sets up the field should be regarded as zero potential energy:

$$P_\infty = 0 \quad (17.7)$$

In that case we obtain for a charge displaced from point  $B$  to infinity

$$W_{B\infty} = P_B - P_\infty = P_B \quad (17.7a)$$

This condition means that the potential energy of a charge at some point of the field is numerically equal to the work performed by the forces of the field in displacing the charge from that point to infinity. Hence, if the field is made by a positive charge the potential energy of another positive charge at some point of this field will be positive, while if the field is established by a negative charge, the potential energy of a positive charge in this field will be negative. For a negative charge in an electric field the reverse will be true. (Why?)

When the field is made by several charges, the potential energy of a charge  $q$  at some point  $B$  of the field is the algebraic sum of its energies in the fields (at point  $B$ ) of all the individual charges. We recall that the electric field strengths of the individual charges at every point of the space are also added up (but like vectors are). Hence, if the fields of several charges exist simultaneously in space, they are simply superimposed on each other. This property of fields is termed *superposition*.

We note in addition that in electrical engineering zero potential energy is often attributed to charges connected with the Earth. In this case the potential energy of a charge at some point of the field is numerically equal to the work performed by the forces of the field in moving this charge from that point to the surface of the Earth.

### 17-6 Electric Potential and Potential Difference

It was established in the preceding section that the potential energy of an electric charge depends on its position in the electric field. Therefore it is now expedient to introduce an energy characteristic for the points of an electric field.

Since the force acting on a charge is directly proportional to its magnitude, the work of the field in displacing the charge should also be proportional to it. For this reason the potential energy of a charge at an arbitrary point  $B$  of the electric field must also be directly proportional to the magnitude of the charge:

$$P_B = \varphi_B q \quad (17.8)$$

The proportionality factor  $\varphi_B$  remains constant with time for every point of the field and can serve as an energy characteristic of the field at this point.

The term for the energy characteristic  $\varphi$  of an electric field at some specified point is field's *potential* at this point.

The measure for the potential is the potential energy of a positive unit charge at this point:

$$\varphi_B = P_B/q \quad (17.8a)$$

It follows from Section 17-5 that the potential at a point of an electric field is numerically equal to the work performed by the forces of the field in displacing a positive unit charge from this point to infinity. (Consider the conditions under which this statement is valid.)

The potential at a specified point can be calculated theoretically. It is determined by the sign, magnitude and arrangement of charges producing the electric field and by the medium. Such calculations are too complex to be presented here. We will simply write out the formula for the potential of the field of a point charge  $q$  obtained with the aid of such calculations.

It can be demonstrated that in a medium with permittivity  $\epsilon_m$  the potential of a point charge  $q$  at point  $I$ , a distance  $r_1$  away from the charge (Fig. 17.11), is

$$\varphi_1 = \frac{q}{4\pi\epsilon_m r_1} \quad (17.9)$$

Note that the same formula is used to calculate the potential of a field produced by a charge  $q$  uniformly distributed over the surface area of a sphere, at all points outside the sphere  $r_1$ , this being the distance from the centre of the sphere to point  $I$ . (Work out when the potential calculated with the aid of formula (17.9) will be positive and when it will be negative.)

It should be noted that the potential of a field produced by a positive charge decreases with the distance from the charge, while the potential of a field produced by a negative charge increases with the distance from it. Since the potential is a scalar quantity, the potential of a field produced by several charges at any point is equal to the algebraic sum of the potentials produced by the individual charges at that point.

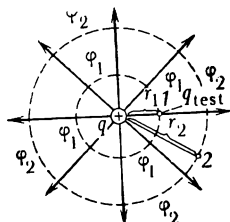
The work of the forces of a field can be expressed with the aid of *potential difference*. We recall that the work in displacing a charge from point 1 to point 2 (see Fig. 17.11) is expressed by formula (17.6a):

$$W_{12} = -\Delta P_{21} = -(P_2 - P_1)$$

Substituting the expression for  $P$  from formula (17.8) we obtain

$$W_{12} = -(\varphi_2 q_{\text{test}} - \varphi_1 q_{\text{test}}) = -q_{\text{test}} (\varphi_2 - \varphi_1) = -q_{\text{test}} \Delta\varphi$$

Fig. 17.11 Electric potential is equal at all points equidistant from point charge.



But this can also be written in the form

$$W_{12} = q_{\text{test}}(\varphi_1 - \varphi_2)$$

The potential difference  $(\varphi_1 - \varphi_2)$  is termed the *voltage* between points 1 and 2 and is denoted  $U_{12}$ . Hence

$$W_{12} = q_{\text{test}}U_{12}$$

Discarding the indices we obtain

$$W = qU \quad (17.10)$$

Hence, the work of the field in displacing the charge  $q$  from one point of the field to another is directly proportional to the voltage between these points.

We deduce a unit of voltage from (17.10):

$$U = \frac{W}{q}, \quad U = \frac{1 \text{ J}}{1 \text{ C}} = 1 \frac{\text{J}}{\text{C}} = 1 \frac{\text{kg} \cdot \text{m}^2}{\text{s}^2 \cdot \text{A}} = 1 \text{ V (volt)}$$

The unit for measuring voltage in the SI system is the *volt*. One volt is the term for a voltage (potential difference) between two points of a field which requires the field to perform work of 1 joule in order to move a charge of 1 coulomb from one of those points to another. Note that in practice charges always move between definite points in the field. Therefore in most cases it is important to know not the potentials at these points, but the voltages between the different points.\*

It follows from (17.9) that the potential  $\varphi_1$  is the same at all points of the field at a distance  $r_1$  from the point charge  $q$  (see Fig. 17.11). All these points lie on the surface of a sphere of radius  $r_1$  with its centre at the location of the point charge  $q$ .

A surface whose all points are of the same potential is termed *equipotential* (from the Latin *aequi* for equal). The cross sections of such surfaces with potentials  $\varphi_1$  and  $\varphi_2$  are shown as circles in Fig. 17.11. For an equipotential surface

$$\varphi = \text{constant} \quad (17.11)$$

It is an established fact that the lines of strength of an electric field are always normal to equipotential surfaces. This means that the work of the field in displacing charges along an equipotential surface is always zero. (Demonstrate that this follows directly from formula (17.10).)

\* Note that in the Gaussian system formula (17.9) assumes the form

$$\varphi = \frac{q}{\epsilon r} \quad (17.9a)$$

the unit of voltage in this system, statvolt (statV) being equal to 300 V, that is, 1 statV = 300 V.

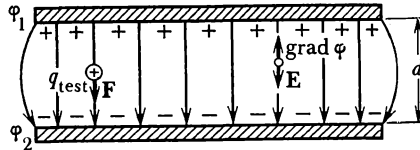
Since the work of a field in displacing the charge  $q$  is entirely determined by the potential difference between the starting and terminal points of its path, the work in moving the charge from one equipotential surface to another (their potentials being  $\varphi_1$  and  $\varphi_2$ ) is independent of path and equal to  $W = q(\varphi_1 - \varphi_2)$ .

One should keep in mind that a field always moves positive charges from points of greater potential to points of smaller potential, the reverse being true for negative charges.

### 17-7 Relation Between Electric Field Strength and Voltage

There is a definite relationship between the field strength and the potential difference in an homogeneous field. To establish this relationship let us recall the expression for the

Fig. 17.12 Greater potential difference between plates carrying unlike charges effects greater field strength between them.



work of a homogeneous field in displacing the charge  $q_{\text{test}}$  in such a field (Fig. 17.12).

Suppose the voltage between the plates is

$$U = \varphi_1 - \varphi_2$$

Now, in moving the charge  $q_{\text{test}}$  from one plate to the other the field will perform work

$$W = q_{\text{test}}U \quad (17.12)$$

The same work can be expressed as a product of the electric force  $F$  and the path  $d$ :

$$W = Fd = q_{\text{test}}Ed \quad (17.13)$$

Equating the right-hand sides of formulae (17.12) and (17.13), we obtain

$$q_{\text{test}}Ed = q_{\text{test}}U$$

whence

$$E = U/d = (\varphi_1 - \varphi_2)/d \quad (17.14)$$

or

$$E = -(\varphi_2 - \varphi_1)/d = -\Delta\varphi/d \quad (17.14a)$$

Therefore the strength of a homogeneous field is numerically equal to the variation of the potential per unit length of the line of strength. In accordance with formula (17.14) the term for the measuring unit of field strength in the SI system is *volt per metre* (V/m). Indeed,

$$E = U/d, \quad E = 1 \text{ V}/1 \text{ m} = 1 \text{ V/m}$$

(Consider how you would demonstrate that  $1 \text{ V/m} = 1 \text{ N/C}$  (see Section 17-2).)

The variation of potential per unit length of a line of strength of a homogeneous electric field is termed *potential gradient* and denoted  $\text{grad } \varphi$ . If one regards the gradient as a vector pointing in the direction of the maximum increase in the potential, it can be easily established that at every point the field strength vectors and the field gradient are opposite in direction (see Fig. 17.12). It can be rigorously proved that these vectors are always equal in magnitude and opposite in direction at any point of a field (including non-homogeneous fields\*).

It follows from formula (17.14) that the strength is greater in those parts of the field where the potential variation per unit length of a line of strength is at its greatest. This may be expressed briefly: the field is stronger in places where the potential changes more quickly.

### 17-8 A Conductor in an Electric Field

It is well known that a property peculiar to all conductors is the presence in them of a large number of mobile charge carriers, i.e. of free electrons or ions.

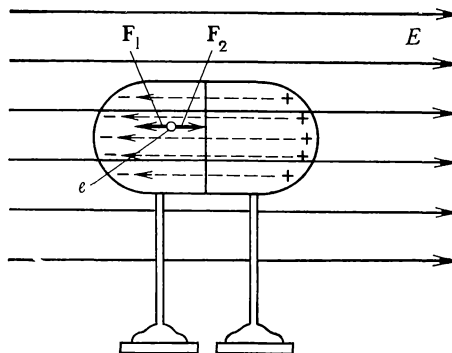
Inside the conductor these charges, in general, move at random. However, in the presence of an electric field in the conductor a directional motion of the charge carriers in the direction of the electric forces is superimposed on their random motion, this directional motion always tending to reduce the field in the conductor. Since the number of mobile charge carriers in a conductor is very large ( $1 \text{ cm}^3$  of a metal contains about  $10^{23}$  free electrons), the motion of the charge carriers acted upon by an external electric field continues until the field inside the conductor vanishes completely. Let us examine this process in more detail.

\* In a nonhomogeneous field  $\text{grad } \varphi = \Delta\varphi/\Delta r$  is defined as the potential variation  $\Delta\varphi$  along such a short segment  $\Delta r$  of a line of strength that the field along it may be considered as homogeneous.



Suppose a metal conductor made up of two parts tightly pressed together is placed in an electric field  $E$  (Fig. 17.13). The field in this conductor acts on the free electrons with a force  $F_1$  from right to left, that is, in the direction opposite to that of the field strength vector. (Explain why.) The

**Fig. 17.13** Conductor in electric field. Solid lines are lines of strength of external field, dashed lines are lines of strength of internal field (field of displaced charges).



displacement of the electrons acted upon by those forces causes an excess positive charge to be accumulated on the right end of the conductor, the left end carrying an excess electron charge. The result is an internal field shown in Fig. 17.13 by dashed lines. Inside the conductor this field is opposed to the external field and acts on every free electron in the conductor with a force  $F_2$  directed to the right.

Initially the force  $F_1$  exceeds the force  $F_2$  (in magnitude) and their resultant points to the left. Accordingly, the electrons inside the conductor continue to move to the left, increasing the internal field. When a sufficient number of free electrons (it is always a negligible fraction of the total) is accumulated on the left end of the conductor, the force  $F_2$  becomes equal to the force  $F_1$  (in magnitude) and their resultant vanishes. After that moment the free electrons inside the conductor will be engaged only in random motion. And this means that the field strength inside the conductor has become zero, that is, the field inside the conductor has vanished.

In short, when a conductor is placed in an external electric field, its electrification is such that a positive charge is accumulated on one end of the conductor and a negative charge of equal magnitude on the other. Such electrification is called *electrostatic induction*. Note that this process involves only the redistribution of the intrinsic charges of the conductor. Therefore, if such a conductor is withdrawn from the field,

its positive and negative charges will again be distributed uniformly over the entire conductor and it will again be neutral at any point.

It can be easily established that there are actually equal charges of opposite sign on the opposite ends of a conductor electrified by induction. Divide the conductor in two (see Fig. 17.13) and withdraw both parts from the field. Connecting each of them to a separate electroscope, we see that they are charged. (Work out how one could prove that the charges are of different sign.) If we join both parts to form a single conductor, we find that the charges have been neutralized. This means that before they were joined the charges were equal.

The time it takes to electrify a conductor by induction is so small that the equilibrium of the charges on the conductor is established practically instantaneously. In equilibrium the field strength, and with it the potential gradient, inside the conductor become zero. Then for any two points inside the conductor the relation will hold

$$\varphi_1 - \varphi_2 = 0, \text{ that is, } \varphi_1 = \varphi_2$$

This means that when the charges on the conductor are in equilibrium, the potential is the same at all points of the conductor. This is also true for a conductor electrified by contact with a charged body. Take a conducting sphere and place a charge  $q$  at a point  $M$  on its surface (Fig. 17.14). The result will be a short-duration excess charge at point  $M$  and a field in the conductor. Acted upon by this field the charge will be uniformly distributed over the entire surface of the sphere and the field will vanish inside the conductor.

Thus, no matter what method is used to electrify a conductor in conditions of charge equilibrium, there will be no field inside the conductor and the potential of all its points will be the same (both inside the conductor and on its surface). At the same time, the field outside the conductor, naturally, continues to exist and its lines of strength are normal (perpendicular) to the conductor's surface. This will be clear from the following considerations. If a line of strength at some point was inclined to the conductor's surface (Fig. 17.15), the force  $\mathbf{F}$  acting on the charge  $q$  at this point of the surface would be resolved into components  $\mathbf{F}_1$  and  $\mathbf{F}_2$ . In that case the charges acted upon by the force  $\mathbf{F}_2$  would move across the conductor's surface. But this is impossible in conditions of charge equilibrium. Consequently, in conditions of charge equilibrium its surface is an *equipotential surface*.

Fig. 17.14 Charge placed at  $M$  is distributed over entire surface of conducting sphere.

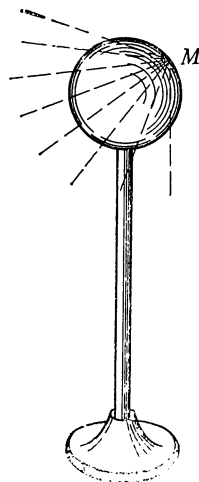


Fig. 17.15 Charges on a conducting surface cannot be in equilibrium if lines of strength are not normal to surface.

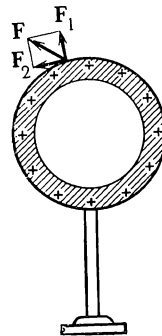
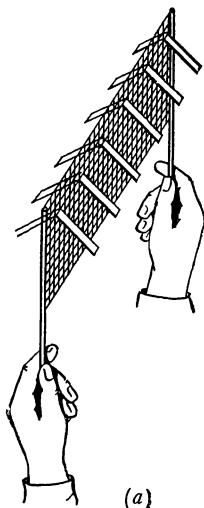
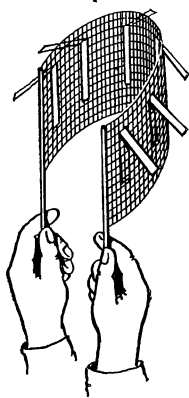


Fig. 17.16 (a) Charges are distributed over both surfaces of metal gauze; (b) all charges went over to external surface of gauze.



(a)



(b)

If there is no field inside the conductor, the body charge density (the amount of electricity per unit volume) in it must be zero everywhere. Indeed, should there be a charge  $q$  in some small volume\* of the conductor it would be surrounded by an electric field.

The theory of the electric field proves that in equilibrium the entire excess charge of an electrified conductor is concentrated on its surface. This means that the entire internal part of the conductor can be eliminated without changing the distribution of its surface charge. For instance, the electric fields around two identically electrified isolated metal spheres, one of which is solid and the other a shell, are identical. The first to prove this in experiment was Michael Faraday.

Hence, if a spherical shell is placed in an electric field or electrified by contact with a charged body, there will be no field inside the shell in conditions of charge equilibrium. This is the principle underlying *electrostatic shielding*. If an instrument is placed inside a metal casing, its operation and readings will not be influenced by the existence of, or variations, in external electric fields, because the external electric fields will not be able to penetrate into the casing.

We will now examine the distribution of charges on the external surface of a conductor. Take a piece of metal gauze with paper strips glued to it and attach it to two insulating handles (Fig. 17.16). If the gauze is charged and then pulled straight (Fig. 17.16a), the strips on both sides will move away from the gauze. If, on the other hand, the gauze is bent to form a ring, only the strips outside will be deflected (Fig. 17.16b). By bending the gauze to form different patterns one can establish that the charges are concentrated only on the external side of the surface, their concentration being greater in places of greater curvature (of smaller radius of curvature).

Hence, the charge distribution is uniform only on a spherical conducting surface. For a surface of arbitrary shape the surface charge density  $\sigma$  and, consequently, the field strength close to the conductor's surface is greater in places of greater curvature. The charge density is especially high on protrusions and points of a conductor (Fig. 17.17). We can prove this by touching with a probe first the conductor at different points and then the electroscope.

\* The term *small volume* implies here a *microscopic volume* that still contains a great number of molecules.

An electrified conductor with protrusions or a point soon loses its charge. Therefore a conductor which has to store a charge for a long time should have no protrusions. (Why does the rod of an electroscope have a ball on its end?)

### 17-9 The Electrometer

The instrument used for measuring the potential of a charged conductor with respect to the Earth or some other charged conductor is termed the *electrostatic voltmeter*, or *electrometer*.

The simplest type of electrometer consists of an aluminum pointer attached to a metal rod and able to rotate about a horizontal axis (Fig. 17.18). The pointer and the rod are inside a metal casing and are well insulated from it with the aid of a plug made of a good dielectric. A small slot is made in the casing through which the scale and the tip of the pointer are visible.

Since the external field does not enter the electrometer, its readings depend only on the potential difference between the casing and the rod. In order to measure the potential of a charged conductor with respect to the Earth one has to connect the conductor to the electrometer rod by means of a metal wire and to ground the electrometer casing, that is, to connect it to the Earth by means of a wire. Then the potential of the conductor will be determined from the position of the pointer on the electrometer scale, usually calibrated in volts. At this point it should be noted that the Earth is a good conductor of electricity and that the charges on it are practically in a state of equilibrium. Accordingly, the potential of all the points on the Earth can, with sufficient accuracy, be assumed to be equal. Hence, in the absence of electric current any conductor connected to the Earth has the same potential as the Earth, that is, zero.

If it is required to measure the potential difference between any two conductors, the electrometer casing is disconnected from the Earth, then one of the conductors is connected by means of a wire to the electrometer casing and the other to the rod. The potential difference is taken from the reading on the electrometer's scale. (Work out what the readings on the electrometer will be if both conductors are connected by means of a wire.)

### 17-10 A Dielectric in an Electric Field

Let us look into the processes at work in a dielectric placed in an electric field. It is well known that there are no free

Fig. 17.17 Distribution of charges over irregular shaped surface is nonuniform.

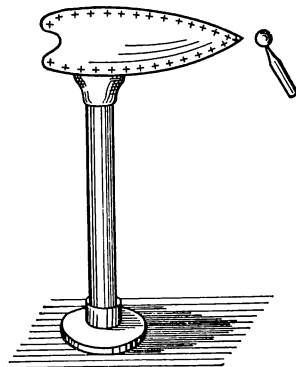


Fig. 17.18 Electrometer.

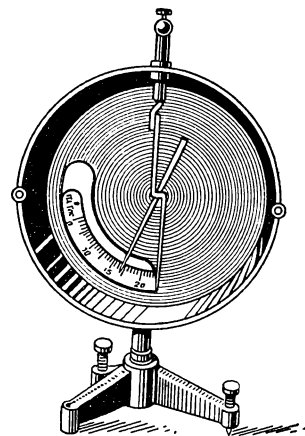


Fig. 17.19 In external field, electron cloud is displaced with respect to nucleus and centres of positive and negative charges no longer coincide.

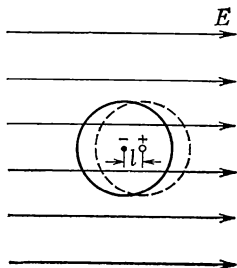


Fig. 17.20 Electric field of dipole reduces external electric field.

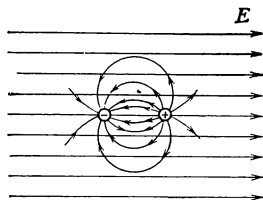
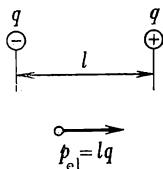


Fig. 17.21 Electric moment of dipole.



charge carriers in a dielectric. All the electric charges in a dielectric are part of its molecules and can be displaced only over very short distances inside the bounds of a molecule or atom.

Since a dielectric weakens the force of charge interaction, that is, weakens the electric field (see Section 16-7), it can be assumed that the charge displacement inside the dielectric's molecules does, in fact, take place. Let us examine the mechanism of this phenomenon.

First, imagine an atom with a nucleus about  $10^{-15}$  m in diameter. Then its electron cloud (as a first approximation assume it to be spherical) will have a radius of about  $10^{-10}$  m. Comparing the dimensions of the nucleus with those of the electron cloud, we see that the atomic nucleus can easily be regarded as a point in the centre of the cloud. If such an atom is placed in an electric field of strength  $E$ , the cloud will be displaced in the direction opposite to  $E$  by a distance  $l$  with respect to the nucleus (Fig. 17.19).

Since the mass of the nucleus is several thousand times greater than that of the electron and since the latter moves in the atom at a very great speed (of the order of  $10^6$  m/s), the nucleus feels only the average force of attraction of the electrons in the atom. Because of this the entire negative charge of the cloud can be assumed to be concentrated in its centre, the entire atom in an electric field being likened to a system of two charges  $q = Ze$ , equal in magnitude and opposite in direction and spaced at a distance  $l$ . Such a system is called a *dipole*. Hence, an atom placed in an electric field becomes an electric dipole, which sets up its own electric field and so weakens the external field in the dielectric (Fig. 17.20).

The product  $p_{el} = lq$  is termed the *electric dipole moment*. The electric dipole moment  $p_{el}$  is a vector directed along  $l$  from the negative charge to the positive (Fig. 17.21) and whose modulus is given by the relation

$$p_{el} = lq \quad (17.15)$$

It is an established fact that the electric dipole moment of molecules due to the displacement of their electron clouds with respect to their nuclei, is directly proportional to  $E$ :

$$p_{el} = \alpha E \quad (17.16)$$

(the term for  $\alpha$  is the *polarizability* of the molecule). Accordingly, the greater the external field strength  $E$  the greater the electric dipole moments in the dielectric, all the electric moment vectors of its molecules thus pointing in the direction

of  $\mathbf{E}$ . Such a dielectric is said to be *polarized* and its dipoles are said to be *soft*, since their length  $l$  depends on  $\mathbf{E}$ .

The polarization of a dielectric involving the displacement of the electron clouds in molecules with respect to their nuclei is termed *electronic polarization*. It can be observed in all dielectrics, the interesting point about it being its independence of temperature.

A molecule which has no centre of symmetry possesses an intrinsic electric moment even in the absence of an electric field in the dielectric (Fig. 17.22). To simplify the discussion the atoms in such a molecule can be regarded as forming a rigid bond, so that the molecule's electric dipole moment is independent of the external field in the dielectric. Such dipoles are termed *rigid*. As examples of natural dipoles one can cite, for instance, the molecules of water, in which the atoms are arranged as in Fig. 17.22*b* (the OH bonds make an angle of  $\sim 105^\circ$  to each other).

If there is an intrinsic electric field surrounding every natural dipole in a dielectric, the question is: Is it possible for a dielectric to set up its own external electric field, even in the absence of other external fields, in the same way as permanent magnets set up magnetic fields? Such dielectrics do, in fact, exist. They will be discussed below (see Section 17-11). In most dielectrics the molecules are, however, arranged at random, therefore, their electric fields are compensated and there is no electric field outside (or inside) the dielectric. If, on the other hand, such a dielectric is placed in an electric field, there will be a force couple acting on every dipole (Fig. 17.23*a*). Therefore rigid dipoles turn and in strong fields even form chains along the field strength lines (Fig. 17.23*b*). In such a situation the electric fields of the individual dipoles augment each other and the dielectric sets up its own electric field (Fig. 17.23*c*). This phenomenon is termed *orientation*, or *dipole polarization* of a dielectric. It can easily be reasoned that orientation polarization should decrease with a rise in temperature, since the random motion of the dipoles breaks up their regular order in the polarized dielectric.

If a dielectric contains ions, a third type of polarization can be observed. Acted upon by an external field the positive ions of a dielectric are displaced in the direction of the strength vector, the displacement of the negative ions being in the opposite direction. Such a phenomenon is known as *ionic polarization* of a dielectric.

The vector sum of the electric dipole moments of all the dipoles contained in a unit volume of a dielectric is termed the *electric polarization vector*,  $\mathbf{P}_{el}$ . In an unpolarized dielec-

Fig. 17.22 Two possible configurations of  $A_2B$  molecule: (a) net dipole moment is zero; (b) there is nonzero dipole moment equal to vector sum of dipole moments of individual bonds.

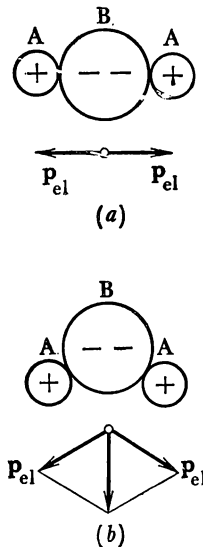
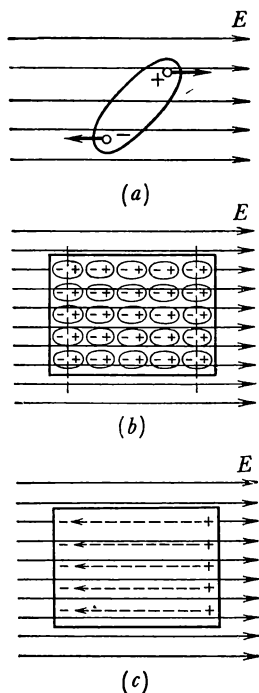


Fig. 17.23 Dielectric in electric field: (a) force couple acts on every dipole; (b) dipoles line up along field strength lines; (c) intrinsic field is set up in dielectric.



tric  $P_{el}$  is zero, while in a polarized one it is directly proportional to  $E$  (if  $E$  is not too great).

It can be seen from Fig. 17.23b that the ends of adjacent dipoles carrying opposite charges must neutralize their action on other charges. Only the charges of the dipole ends terminating on the dielectric's surface remain uncompensated. The negative dipole charges terminate on the surface through which the strength lines of the external field enter the dielectric, and the positive charges terminate on the opposite surface. All the charges on the surface of a polarized dielectric are *bound charges*, that is, they are part of molecules. The term for them is also *polarization charges*. The effect a polarized dielectric has on an external electric field is due entirely to its polarization charges.

The field inside a dielectric set up by its polarization charges is directed against the external field (Fig. 17.23), that is, it tends to weaken the external field but does not eliminate it completely (compare this with a conductor). Another distinction of a dielectric from a conductor is that it is impossible to separate the positive charges from the negative by taking the dielectric apart. The opposite surfaces of a polarized dielectric will always retain charges of opposite sign. This proves that the polarization charges are, in fact, bound, or that they are part of the dipoles.

Note in addition that the modulus of the electric polarization vector is equal to the surface charge density of the dielectric:

$$P_{el} = \sigma_{pol} \quad (17.17)$$

This can easily be proved if the polarized dielectric in the shape of a parallelepiped is regarded as one big dipole\*. It follows from (17.17) that the unit for measuring  $P_{el}$  in the SI system is  $1 \text{ C/m}^2$ .

The weakening of the field in a dielectric by its polarization is used to explain the effect of the dielectric on the force of interaction between electrified bodies. Indeed, if two charges  $q_1$  and  $q_2$  are immersed in a dielectric, it will become polarized and the charges  $q_1$  and  $q_2$  will be surrounded by polarization charges. This is equivalent to the reduction of the charges  $q_1$  and  $q_2$  (Fig. 17.24) and, consequently, to a decrease in the force acting between them. It is now understandable why the force of interaction between charges is at its maximum in a vacuum and why the formula for Coulomb's law includes the permittivity of the medium,  $\epsilon_m$ .

\* Try and do this yourself, taking into account that the electric moment of such a dipole can be expressed both in the form  $ql = \sigma_{pol}Al$  and  $P_{el}V$ , where  $V$  is the volume of the parallelepiped.

Note that the dipoles of a dielectric may break up if the field is high enough. In this case free charge carriers appear in the dielectric. Their motion results in the mechanical destruction of the dielectric. The term for this phenomenon is *electric breakdown* of the dielectric. An electric discharge in the form of lightning during a thunderstorm serves as an example of such a breakdown.

### 17-11 Ferroelectrics

Studies of solid dielectrics containing natural dipole molecules led to the discovery of a group of substances with quite unusual electrical properties termed *ferroelectrics*. It was established that a molecule of a ferroelectric has no centre of symmetry and that it is strongly anisotropic in its properties.

The permittivity of the ferroelectrics proved to be temperature-dependent with maximum values lying inside definite temperature ranges. For instance, the maximum relative permittivity (dielectric constant) of Seignette salt is about 10 000, and that of barium titanate ( $\text{BaTiO}_3$ ) about 7000.

It was also established that the dielectric constant of ferroelectrics is field-dependent in comparatively weak fields (in normal dielectrics such dependence becomes noticeable only at close to breakdown level). This means that the polarization vector of ferroelectrics is not proportional to  $\mathbf{E}$ .

Finally, the electric polarization vector of a ferroelectric depends not only on  $\mathbf{E}$  but on its history as well. Such a dependence is termed *dielectric hysteresis* (from the Greek *hysteresis* for lagging). In an alternating field of high enough strength the  $p$  versus  $E$  plot forms a closed loop termed the *hysteresis loop* (Fig. 17.25). After the external electric field is switched off, the ferroelectric retains a residual polarization (section  $OD$ ).

All these properties of the ferroelectrics are the result of their peculiar internal structure. A ferroelectric is made up of spontaneously polarized regions termed *domains*. In an unpolarized ferroelectric the domains are arranged at random and their polarization vectors are mutually neutralized. Hence they do not produce a field outside the ferroelectric (although a small splinter of such a ferroelectric containing several domains may have perceptible polarization). In an ideal ferroelectric with a rectangular hysteresis loop all domains turn in the direction of an external field acting on the ferroelectric. In real-life ferroelectrics their orientation is obstructed by the thermal motion of the domains and by

Fig. 17.24 Dipoles reduce interaction force between charges introduced into dielectric.

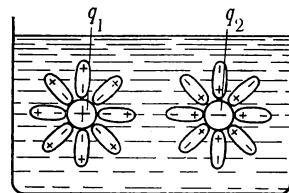
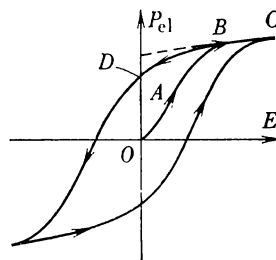


Fig. 17.25 Hysteresis loop of ferroelectric material;  $OD$  is residual polarization.





certain molecular forces causing the hysteresis loop to assume the shape shown in Fig. 17.25. All ferroelectrics must contain electric domains.

It has been established that the domains in ferroelectrics exist only in a definite temperature range, in which they exhibit ferroelectric properties. For instance, in Seignette salt these properties exist only in the temperature range from  $-15^{\circ}\text{C}$  to  $+22.5^{\circ}\text{C}$ .

### 17-12 The Piezoelectric Effect

In the course of research carried out with solid dielectrics it was established that they can be polarized not only by an electric field but also in the process of deformation resulting from mechanical action.

The polarization of a dielectric as a result of mechanical action is termed the *direct piezoelectric effect*. This effect is exhibited by quartz crystals and all ferroelectrics. In order to observe it, one should cut out of a crystal a parallelepiped with its faces precisely oriented in defined directions. Metal electrodes *A* and *B* are deposited on two opposite faces of this parallelepiped and provided with wires for connection into an electrical circuit or instrument (Fig. 17.26).

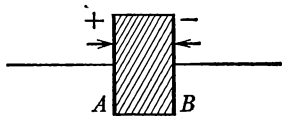
When the parallelepiped is compressed, one of its faces becomes positively charged and the opposite face negatively. It was established that in this case the magnitude of the polarization charge is proportional to the pressure and independent of the dimensions of the parallelepiped. If extension of the parallelepiped is substituted for compression the charges on its faces change signs.

The explanation of the direct piezoelectric effect is as follows. None of the piezoelectric crystals have a centre of symmetry, while all of them are made up of positive and negative ions, the ions of each kind forming interlacing individual sublattices. When the crystal is compressed, the sublattices are displaced with respect to each other, with the result that one surface of the crystal acquires a positive charge and the other a negative charge. When the crystal is extended, the displacement of the sublattices is reversed, together with the charges on the crystal surfaces.

The direct piezoelectric effect is used in microphones, phono pickups, loudspeakers, etc.

Piezoelectric crystals also exhibit the reverse effect. If a plate cut out of a piezoelectric crystal is placed in an electric field by charging the metal electrodes, it will be polarized and deformed, for instance, it may contract. A change of

Fig. 17.26 Polarization of dielectric induced by mechanical action.



sign of the charge results in a change from contraction to extension. The deformation of crystals caused by polarization in an external electric field is termed the *reverse piezoelectric effect*.

If a piezoelectric crystal plate is placed in an alternating electric field, it will vibrate with the frequency of the field. This is the principle underlying the application of the reverse piezoelectric effect for the generation of ultrasonic vibrations (see Section 28-8) in loudspeakers, frequency stabilizers, etc.

### 17-13 Capacitance

Take a conductor insulated from the ground and without moving it start to electrify it. Experiment shows the variation of its charge to be directly proportional to its potential  $\varphi$ :

$$q = C\varphi \quad (17.18)$$

The proportionality factor  $C$  remains constant only if the experimental arrangement is not changed in the process of varying  $\varphi$ , that is, if the conductor itself and material objects close to it remain stationary. If the experiment is carried out with a conductor of different dimensions or shape, or if the position of the former conductor is changed in relation to nearby objects, the numerical value of  $C$  will not remain the same.

The quantity  $C$  characterizing the dependence of the charge of an electrified conductor on its dimensions and shape and on its surroundings is termed the *capacitance* of the conductor. The measure of a conductor's capacitance is the amount of electricity required to raise the potential of this conductor by a single unit.

We deduce a unit for measuring capacitance  $C$ :

$$C = \frac{q}{\varphi}, \quad C = \frac{1 \text{ C}}{1 \text{ V}} = 1 \frac{\text{C}}{\text{V}} = 1 \frac{\text{A} \cdot \text{s}}{\text{kg} \cdot \text{m}^2} = 1 \text{ F (farad)}$$

The unit of capacitance in the SI system is the *farad*. One farad is the capacitance of a conductor which needs charge of 1 C to raise its potential by 1 V. Since the farad is a very large unit (see the end of this section), in practice capacitance is often expressed in microfarads ( $\mu\text{F}$ ) and picofarads (or micro-microfarads) (pF or  $\mu\mu\text{F}$ ),

$$1 \mu\text{F} = 10^{-6} \text{ F}, \quad 1 \text{ pF} = 1 \mu\mu\text{F} = 10^{-12} \text{ F}$$

The capacitance of a conductor of a regular shape can be calculated theoretically. As an example, let us demonstrate how the formula for the capacitance of an isolated conducting sphere of radius  $r$  is obtained. We have from (17.18)

$$C_{sp} = q_{sp} / \varphi_{sp} \quad (17.18a)$$

Since the expression for the potential of a charged sphere is (17.9), that is

$$\varphi_{sp} = \frac{q_{sp}}{4\pi\epsilon_m r_{sp}}$$

by substituting the expression for  $\varphi_{sp}$  into (17.18a) we obtain

$$C_{sp} = \frac{q_{sp} 4\pi\epsilon_m r_{sp}}{q_{sp}}$$

whence

$$C_{sp} = 4\pi\epsilon_m r_{sp} = 4\pi\epsilon_0 \epsilon r_{sp} \quad (17.19)$$

Hence, the capacitance of an isolated conducting sphere is directly proportional to its radius.\*

Using formula (17.19) one can demonstrate that for a sphere in a vacuum to have a capacitance of 1 F its radius must be  $9 \times 10^6$  km. This radius is 23 times the distance from the Earth to the Moon.

It follows from formula (17.19) that

$$\epsilon_m = \frac{C_{sp}}{4\pi r_{sp}}$$

Accordingly, the unit for measuring permittivity in the SI system is the *farad per metre* (F/m):  $1 \text{ C}^2/(\text{N} \cdot \text{m}^2) = 1 \text{ F/m}$  (see Section 16-8).

#### 17-14 Factors That Determine Capacitance

Charges are concentrated exclusively on the external surface of a conductor. Therefore neither the material of the conductor nor its mass are of any importance for its capacitance. It was mentioned above that capacitance depends on the shape and surface area of the conductor. Since a conductor

\* In the Gaussian system formula (17.19) assumes the form

$$C_{sp} = \epsilon r_{sp} \quad (17.19a)$$

The unit for measuring capacitance in the Gaussian system is the *centimetre* and this follows from (17.19a). Note that  $1 \text{ F} = 9 \times 10^{11} \text{ cm}$ . The unit is also called the *statfarad* (statF).

is liable to be electrified by induction, its capacitance is influenced by other conductors in its vicinity and by the medium they are in. Let us demonstrate this in an experiment.

Take two metal disks mounted on dielectric supports. Connect the disk  $A$  to an electrometer with its casing grounded (Fig. 17.27) and move disk  $B$  away from disk  $A$ . Electrify disk  $A$  by imparting to it a charge  $q$  which remains constant.

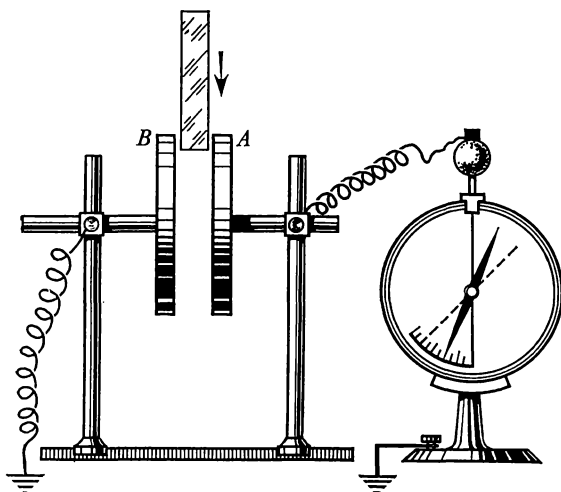


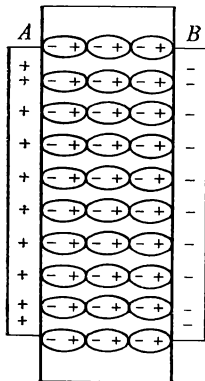
Fig. 17.27 Capacitance of metal disks  $A$  and  $B$  depends on their separation and the dielectric between them.

Look at the readings on the electrometer and record the potential of  $A$ . Bring the disk  $B$  nearer to disk  $A$  and observe the position of the pointer. The potential  $\phi$  of disk  $A$  will be seen to diminish.

The decrease in the potential  $\phi$  will be still more rapid if disk  $B$  is grounded. If one bears in mind that the charge  $q$  on disk  $A$  remains constant in the process, one can, using formula (17.18), conclude that the capacitance of the system of disks increases. Substituting other dielectrics for the air in the space between the disks, one will again observe the capacitance of the system of disks to increase.

The explanation of these results is as follows. When disk  $B$  is placed in the field of disk  $A$ , it is electrified by induction and sets up its own field. If disk  $B$  is grounded, only the charges of opposite sign to those of disk  $A$  will remain on it. This augments the field of disk  $B$  and, in turn, reduces the potential of disk  $A$ . When we insert a dielectric (for instance, glass) between the disks, it is polarized. The polarization charges close to the surface of  $A$  bind some of its charges.

Fig. 17.28 Dielectric placed between disks  $A$  and  $B$  increases capacitance.



Therefore the surface density of charges on disk  $A$  can now be assumed to be  $\sigma - \sigma_{\text{pol}}$  instead of  $\sigma$  (Fig. 17.28). And this means that the capacitance of the disk has increased.

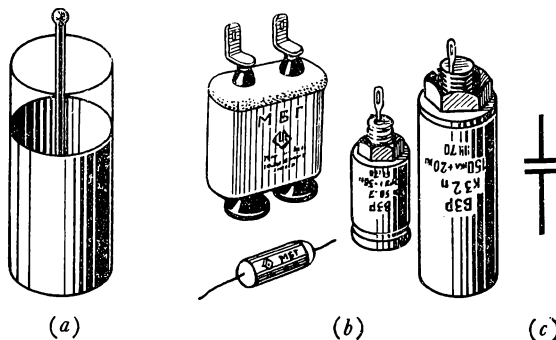
### 17-15 Capacitors

Radio receivers, television sets, tape recorders and other types of electronic equipment employ *capacitors*—devices for the storage of electric charges and energy whose capacitance is independent of external conditions, that is, remains constant during operation.

To fulfil its functions the capacitor must be able to store accumulated charges and energy for an appreciable time. To obtain a definite capacitance one can conveniently take two conductors and arrange them as close to each other as possible and place a dielectric between them. The dielectric between the conductors plays a two-fold role: firstly, it increases the capacitance and, secondly, prevents the neutralization of the charges, that is, prevents them from jumping from one conductor to the other. For this reason its permittivity and electrical breakdown strength should be high.

The two conductors on which charges are accumulated are termed *capacitor plates*. In order to keep the capacitance constant and independent of surrounding bodies the entire electric field should be contained between the plates. For this reason the distance between the plates should be small

Fig. 17.29 (a) Leyden jar; (b) some types of capacitors; (c) schematic symbol for capacitor.



as compared to their linear dimensions. To protect the capacitor from external influences it is housed in a shell.

The accumulation of charges on the capacitor electrodes is termed *charging*. The neutralization of a capacitor's charge by connecting its electrodes across a conductor is termed *discharge*. The amount of electricity  $q$  which crosses from one of the capacitor's plates to the other in the process of its

discharge is termed its *charge*. It is directly proportional to the voltage  $U$  across the capacitor's electrodes. Then the formula for the capacitance of a capacitor is

$$C = q/U \quad (17.20)$$

To charge a capacitor one connects its plates to the terminals of an electric power source, for instance to a battery. Each production model of a capacitor is designed for a defined safe voltage. If the voltage across the capacitor exceeds the peak voltage it will break down. Such a capacitor can no longer be used.

A capacitor whose plates are flat surfaces is called a *plane-parallel capacitor*. Let  $A$  be the surface area of one plate,  $d$  the thickness of the dielectric, and  $\epsilon_m$  its permittivity. The formula for the capacitance of a plane-parallel capacitor is therefore\*

$$C = \epsilon_m A/d \quad (17.21)$$

The external appearance of some capacitors is illustrated in Fig. 17.29.

In a charged capacitor the unlike charges experiencing mutual attraction are concentrated on the internal surfaces of its plates. When the plates are moved, the effective area of the capacitor and with it its capacitance are changed.

Fig. 17.30 When capacitor plates are moved, charges accumulate on surfaces facing each other.

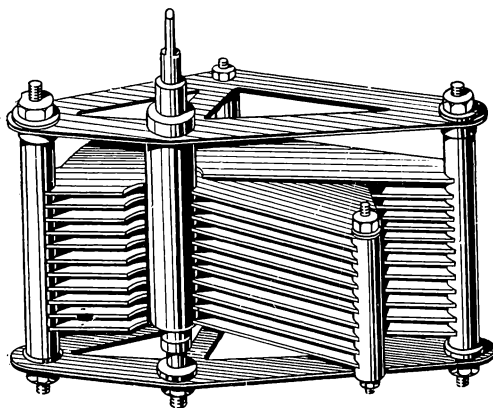


Fig. 17.31 Variable capacitor.

With the reduction in the effective area of the capacitor the charges are concentrated on the parts of the plates that face each other (Fig. 17.30). This property is used in the *variable capacitor* (Fig. 17.31), which is employed, for instance, for

\* In the Gaussian system this formula takes the form

$$C = \frac{\epsilon A}{4\pi d}$$

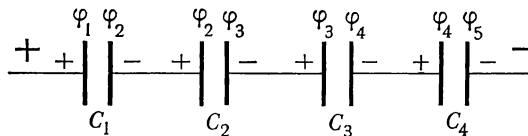
tuning radio receivers. In this capacitor a group of plates carrying charges of one sign (the rotor) is displaced in relation to the other group (the stator).

### 17-16 Combinations of Capacitors in Parallel and in Series

Quite frequently in order to obtain the required capacitance one has to connect capacitors into a group.

The term *in series* applies to a connection of capacitors in which the plate of the preceding capacitor carrying a negative charge is connected with the plate of the following capacitor carrying a positive charge (Fig. 17.32). The plates of

Fig. 17.32 Capacitors in series: capacitors carry equal charges but voltages across them are different.



all the capacitors connected in series carry charges of identical magnitude,  $q$ . (Explain why.) Since the charges are in a state of equilibrium, the potentials of the plates connected by wires will be equal.

Taking these points into account let us deduce a formula for computing the capacitance of capacitors connected in series. It can be seen from Fig. 17.32 that the voltage across the whole set,  $U$ , is equal to the sum of the voltages across the individual capacitors connected in series. Indeed,

$$(\varphi_1 - \varphi_2) + (\varphi_2 - \varphi_3) + \dots + (\varphi_{n-1} - \varphi_n) = \varphi_1 - \varphi_n$$

or

$$U_1 + U_2 + \dots + U_n = U$$

Making use of the relation  $q = CU$ , we obtain

$$\frac{q}{C_1} + \frac{q}{C_2} + \dots + \frac{q}{C_n} = \frac{q}{C}$$

Cancelling out  $q$  we have

$$\frac{1}{C} = \frac{1}{C_1} + \frac{1}{C_2} + \dots + \frac{1}{C_n} \quad (17.22)$$

It follows from (17.22) that in the case of an in series connection the total capacitance turns out to be less than the smallest of the capacitances of the individual capacitors.

The term *in parallel* applies to a connection in which all plates carrying a positive charge are connected to a single

wire, while the plates carrying the negative charge are all connected to another wire (Fig. 17.33). In this case all the plates with charges of like sign are at the same potential and the total charge of the battery  $q$  is the sum of the charges on the individual capacitors:

$$q = q_1 + q_2 + \dots + q_n$$

Since the voltage across all the capacitors is the same  $U_1 = U_2 = \dots = U_n = U$  we have

$$CU = C_1U + C_2U + \dots + C_nU$$

Cancelling out  $U$ , we obtain the formula for calculating  $C$

$$C = C_1 + C_2 + \dots + C_n \quad (17.23)$$

It follows from (17.23) that in the case of an in parallel connection of capacitors the resulting capacitance exceeds the greatest capacitance of the individual capacitors.

In the manufacture of capacitors a method of in parallel connection shown in Fig. 17.34 is used. This method saves space and materials since the charges are concentrated on both surfaces of the plates (excluding the two outside plates). The capacitor in Fig. 17.34 has six capacitors connected in parallel and seven plates. Therefore in this case the number of capacitors connected in parallel is one unit less than the number of metal plates  $n$ :

$$C = \frac{\epsilon_m A}{d} (n-1)^* \quad (17.24)$$

### 17-17 The Energy of a Charged Capacitor

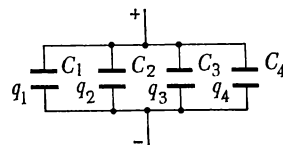
The work of the forces of an electric field in displacing a charge  $q$  from point to point in the field is  $qU$  if the potential difference between the points remains constant in the process. However, when a capacitor is charged, the voltage across its plates increases from zero to  $U$ . Therefore, when one calculates the work of the field one should use the mean value of the voltage. Hence

$$qU_{\text{mean}} = q \frac{U+0}{2} = \frac{qU}{2}$$

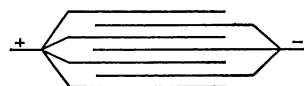
\* In the Gaussian system this formula is written in the form

$$C = \frac{\epsilon A}{4\pi d} (n-1)$$

**Fig. 17.33** Capacitors in parallel: the voltages across the capacitors are equal but their charges are different.



**Fig. 17.34** Diagram of multi-plate capacitor.





Since this work is spent to increase the energy of the charged capacitor, we will use the symbol  $W$  for both quantities. Therefore the expression for the energy of a charged capacitor is

$$W = qU/2 \quad (17.25)$$

Since  $q = CU$ , we obtain another formula for the capacitor's energy

$$W = CU^2/2 \quad (17.25a)$$

The same formulae can be used to compute the energy of a conductor charged with respect to the ground. In this case the voltage is determined from the readings on an electrometer. The student is entitled to ask at this point whether the energy  $W$  is the energy of the charges on the capacitor's plates or whether it belongs to the field between the plates. To clarify the subject let us do the following experiment.

Take a sectional capacitor made of a glass dielectric with copper plates provided with insulating handles. Place the assembled capacitor on a rubber mat and charge it. Next detach the plates from the glass dielectric (take the capacitor apart) and discharge them. After that assemble the capacitor again and short-circuit it with a discharge (an insulated bent metal rod with balls on both ends). A spark will appear between the discharger and the plates. This means that the energy of the electric field transformed into energy of residual polarization, remained stored in the glass dielectric.

In the case of a vacuum such direct experiments are impossible, but the transport of light and radio waves through space prove that a vacuum, too, can contain the energy of an electromagnetic field.

The *energy density*  $w$  of a homogeneous electric field is the term used for the quantity which measures the energy contained in a unit volume of space where such a field exists:

$$w = W/V \quad (17.26)$$

We examine the factors influencing the energy density of a homogeneous electric field of a plane capacitor. Substituting into (17.25a) the value of  $C$  from (17.21), we obtain

$$W = \frac{CU^2}{2} = \frac{\epsilon_m AU^2}{2d}$$

Multiplying both the numerator and the denominator of the right-hand side by  $d$ , we obtain

$$W = \frac{\epsilon_m}{2} \frac{U^2}{d^2} Ad$$

Since  $Ad = V$  and  $E = U/d$ , we obtain  $W = \epsilon_m E^2 V/2$ , whence

$$\frac{W}{V} = w = \frac{\epsilon_m E^2}{2} \quad (17.27)$$

The energy density of an electric field is directly proportional to the square of the strength of the field.

### 17-18 Millikan's Experiment

One of the first experiments which enabled the magnitude of the elementary charge to be determined was carried out in 1917 by the American scientist Robert A. Millikan (1868-1953). The essence of the experiment was as follows.

A tiny droplet of oil of known mass  $m$  was introduced into a homogeneous electric field of a plane-parallel capacitor. The field strength was computed from the formula  $E = U/d$  valid for such a capacitor, the required value of  $E$  being obtained by varying  $U$ . Millikan charged the droplet by ionizing its molecules with radiation and adjusted  $E$  so that the drop remained suspended between the plates (Fig. 17.35). In this case the electric force  $F_{el}$  acting on the droplet is equal (in magnitude) to the force of gravity  $F_{grav}$  acting on it, that is,  $F_{el} = F_{grav}$ , or  $qU/d = mg$ , whence

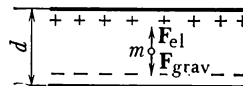
$$|q| = mgd/U$$

Substituting the numbers obtained from the experiment into this formula, Millikan determined the magnitude of the charge,  $q$ . Repeating this experiment a great number of times he established that, with account taken of possible errors, the charge of the droplet obtained in all experiments may be regarded as some integral multiple of some number. Millikan accepted this number as the magnitude of the *elementary charge*. Analyzing the results obtained, he determined the following values for the positive and negative elementary charges:

$$e_+ = 1.60 \times 10^{-19} \text{ C} \quad \text{and} \quad e_- = -1.60 \times 10^{-19} \text{ C}$$

Actually, it was not so simple to retain a drop in the suspended state, so Millikan's experimental arrangement and the formula for computing the results were much more complex.

**Fig. 17.35** Simplified schematic representation of Millikan's experiment for establishing elementary charge.



## 18

## Electric Current in Metals. Direct-Current Circuits

### 18-1 Charge Carriers and Electric Current

When a conductor is placed in an electric field, the mobile charge carriers in the conductor are set in motion by the action of the field, with the result that the potentials of all the points of the conductor are equalized (see Section 17-8). However, if by some means or other the potentials of any two points of the conductor are kept different, an electric field will be maintained in the conductor. This field will be a constant cause of motion of the charges, the positive charges moving from points of greater potential to points of smaller potential, and the negative charges moving in the opposite direction.

The term for the directional motion of charge carriers in a conductor caused by the action of the forces of an electric field is *electric current*. The mobile charge carriers in all conductors are free electrons or ions. Note that the motion of a separate charge carrier in a conductor is sometimes called *microcurrent*, the term for the current of a great number of carriers being *macrocurrent*.

Moving in a vacuum, free charge carriers (of a single sign) do not meet with any resistance and acquire kinetic energy equal to the work of the accelerating field.

When a current flows in a substance, the moving carriers meet with resistance because they interact with other charges and particles constituting the substance. (For instance, electrons in a metal lose energy gained in the electric field in collisions with the ions of the lattice.) Such collisions result in the intensification of random motion, that is, in an increase in the temperature of the substance. Hence, a current flowing in a substance always increases its internal energy.

### 18-2 Current and Current Density

Charges moving in a vacuum are accelerated all the time they are in the accelerating field. The situation in solids and in liquids is quite different.

Just like a train moving at a constant speed when traction is equal to external resistance, a charge carrier moves with constant directional speed when the accelerating force acting

on it is equal to the conductor's resistance. In a field of constant strength the mean velocity of directional motion,  $\mathbf{v}$ , of mobile charge carriers is constant and proportional to the field strength,  $\mathbf{E}$ :

$$\mathbf{v} = u\mathbf{E} \quad (18.1)$$

The term for the proportionality factor  $u$ , which expresses the dependence of the directional velocity of charge carriers acted upon by an electric field on the substance of the conductor and on the surrounding medium is *charge carrier mobility*. The measure of mobility is the speed of directional motion of charge carriers in an electric field of unit strength. (Demonstrate that the unit of mobility in the SI system is  $1 \text{ m}^2/(\text{V} \cdot \text{s}) = 1 \text{ A} \cdot \text{s}^2/\text{kg}$ .)

When an electric current flows in a conductor, charge carriers cross any cross-sectional area  $A$  in the conductor just like cars pass an observer standing on the side of a road. The quantity  $I$  characterizing the rate of charge carrier transfer through a section of the conductor is termed *current*. The measure for the current in a conductor is the amount of electricity crossing a section of the conductor per unit time

$$I = q/t \quad (18.2)$$

The term for the quantity  $j$  characterizing the rate of charge transfer through a unit cross section of a conductor is *current density*. In the case of a uniform distribution of the charge carrier flux over the entire section of the conductor, the current density is

$$j = I/A \quad (18.3)$$

When the distribution of the charge carrier flux over the cross section is not uniform, one can always choose such a small part of the cross section,  $\Delta A$ , that the carrier flux in it can be regarded as uniform. Then, if the current flowing through the chosen area  $\Delta A$  is  $\Delta I$ , the current density in this part of the cross section is

$$j = \Delta I/\Delta A \quad (18.3a)$$

Let us now examine the factors that determine current density in a conductor. Suppose that the field strength lines are perpendicular to the conductor's cross section and that the charge of all carriers is positive and equal to  $e_+$ . Let the field strength in the conductor be  $\mathbf{E}$ , the number of mobile charge carriers per unit volume of the conductor be  $n_0$ , and the velocity of their directional motion be  $\mathbf{v}$ . Isolate a cylinder in the conductor with its generatrix paral-

lel to  $\mathbf{E}$  and its cross section equal to  $A$  (Fig. 18.1). The cylinder will contain  $Aln_0$  charge carriers with a total charge  $q = Aln_0e_+$ .

All these carriers will pass through the lower base of the cylinder in a time  $t = l/v$ , the current through the cylinder's cross section  $A$  being  $I = q/t$  and the current density being

$$j = \frac{I}{A} = \frac{q}{At} = \frac{Aln_0e_+}{Al/v} = n_0e_+v$$

Hence

$$j = n_0e_+v \quad (18.4)$$

or in vector form

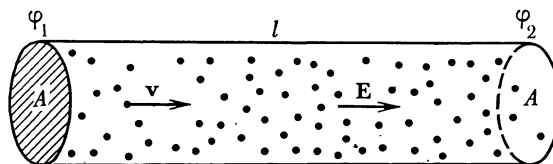
$$\mathbf{j} = n_0e_+\mathbf{v} \quad (18.5)$$

Since  $\mathbf{v} = u\mathbf{E}$ , we obtain

$$\mathbf{j} = un_0e_+\mathbf{E} \quad (18.6)$$

Since the field strength  $\mathbf{E}$  is a vector and  $u$ ,  $n_0$  and  $e_+$  are scalars, the current density  $\mathbf{j}$  coincides in direction

**Fig. 18.1** Calculating current density.



with  $\mathbf{E}$ . It has been agreed to accept the direction of the flux of positive charge carriers as the direction of the current flowing in a conductor. Since  $I = jA$ , we have

$$I = un_0e_+AE \quad (18.7)$$

The field strength in a cylinder can be expressed with the aid of formula (17.14):

$$E = \varphi_1 - \varphi_2/l = U/l$$

We obtain from (18.6) and (18.7)

$$j = \frac{un_0e_+}{l}U \quad (18.6a)$$

$$I = \frac{un_0e_+A}{l}U \quad (18.7a)$$

These formulae can be applied to any conductor of constant cross section and uniform material, with  $l$  being

its total length,  $A$  the cross section, and  $U$  the voltage across its terminals.

The current whose density at all points of the conductor does not change with time is termed *direct current*. It follows from (18.7a) that direct current exists in a conductor only so long as a constant voltage is maintained across its terminals. A current which changes periodically with time is termed *alternating current*. The following applies only to direct current.

It should be stressed once again that the accepted direction of a current is the direction of motion of the positive charges, which coincides with the direction of the field. The negative charges, as we know, are driven by the field in the opposite direction. However, the displacement of negative charges in the conductor in a given direction is equivalent to the displacement of positive charges of equal magnitude in the opposite direction. Therefore the direction of the current due to the negative charge carriers also coincides with that of the electric field. The formulae for current density and current are in this case similar to (18.6a) and (18.7a). If the sign of the charge carriers in these formulae is omitted, they will apply both to positive and to negative charges:

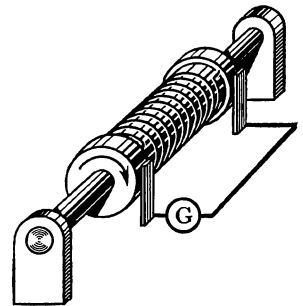
$$j = \frac{un_0e}{l} U \quad (18.6b)$$

$$I = \frac{un_0eA}{l} U \quad (18.7b)$$

The following experiment was undertaken to determine the sign of free charge carriers in a metal. A long metal wire was wound around a cylindrical core and its terminals were connected to two rings (Fig. 18.2) with sliding graphite contacts. The contacts were connected to a galvanometer (an instrument for measuring weak currents). The coil was rotated at great speed and then suddenly braked. At this instant the galvanometer's pointer was deflected, that is, there was a short-lived current. This result may be explained as follows.

When the coil rotated, the charge carriers in the conductor moved with it. When the coil braked suddenly, the charge carriers continued to move, due to inertia, for a period of time, that is, a short-lived current was produced in the conductor. The direction of this current indicated that the charge of mobile charge carriers is negative. More detailed research led to the conclusion that the mobile charge carriers in metals are the electrons.

Fig. 18.2 Rotating coil with sliding contacts.

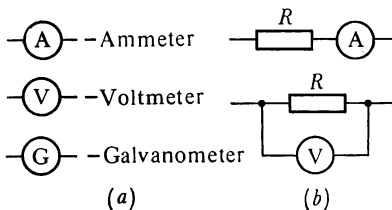


### 18-3 The Ammeter, the Voltmeter and the Galvanometer

In practice it is frequently required to measure the current flowing in a conductor and the voltage across it. The instrument for measuring current is termed *ammeter* and that for voltage, *voltmeter*. These instruments use various current effects as their principles of operation (for instance, liberation of heat in a conductor), but the one most frequently used is the magnetic effect of the current. The design of such instruments will be discussed later.

We recall that in practice both in series and in parallel connections of current-carrying conductors are used. The

**Fig. 18.3** (a) Schematic symbols for ammeter, voltmeter and galvanometer; (b) ammeter is connected in series with conductor, and voltmeter in parallel.



ammeter is always connected in series with the conductor in which the current is being measured, and the voltmeter in parallel with the conductor at the points between which the voltage is being measured (Fig. 18.3).

The instrument that is used to measure weak currents, small voltages and charges is the *galvanometer*. Depending on the type of measurement it can be connected either in series or in parallel with the conductor.

### 18-4 Closed Electric Circuit

One can obtain a short-lived electric current by connecting two charged bodies with different potentials by means of a conductor. The current in the conductor ceases as soon as the potentials of the bodies are equalized. We recall that, in general, current flowing in a conductor tends to weaken the field inside it and equalize the potentials of all the points of the conductor (see Section 17-8).

To obtain a sustained current one has to close the circuit of conductors, so that the charges can circulate in the circuit. Also, the electric field which the current tends to eliminate, should be maintained in the circuit. The field

inside the conductors making up the closed circuit has to be maintained by a power source.

A circuit also includes *consumers* of electric energy, in which the current performs useful work. A circuit includes also *connecting wires*, or *leads*, and a *switch* to connect or disconnect the circuit. Note that the instruments to be connected into a circuit should have *terminals* to which the leads are connected. Accordingly, a simple electric circuit consists of a power source, a consumer, connecting wires and a switch.

### 18-5 Electromotive Force of a Power Source

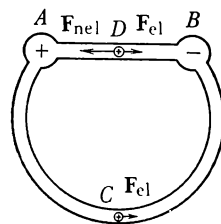
It was demonstrated above that the work of the forces of an electric field in moving a charge in a closed circuit is zero (see Section 17-5). This means that if only the electric forces act on the charges the current can perform no work. Accordingly, the circuit must include at least one section in which nonelectric forces capable of performing work in displacing the charges, or of having work performed by the moving charges against them, act on the charges.

Imagine two charged conductors *A* and *B* (Fig. 18.4). Let the potential of *A* be greater than that of *B*. If these conductors are connected by means of the conductor *ACB*, the positive charges acted upon by the forces of the field  $F_{el}$  will move from *A* to *B* along the conductor *ACB*. However, this motion will cease very quickly since the potentials of *A* and *B* will become equal. To preclude such a situation and to prolong the flow of charge carriers the positive charges have to be, somehow or other, transported back from point *B* to point *A*, for instance, by means of the conductor *BDA*. But such a displacement of the charges cannot take place by itself, since the forces of the electric field drive them in the opposite direction.

Consequently, there must be nonelectric forces  $F_{nel}$  acting on the charge carriers in the conductor *BDA* against the forces of the electric field and exceeding them in magnitude. In that case the carriers in section *ACB* will be driven from *A* to *B* by the electric field and in section *BDA* from *B* to *A* by the nonelectric field. Thus there will be a continuous current flowing in the closed circuit and the potentials of *A* and *B* will not be equalized.

The nonelectric forces will perform work in displacing the charge carriers along the section *BDA* against the electric field and against the resistance of the particles of the substance of which the conductor *BDA* is made. The

Fig. 18.4 In conductor *ACB* the charges move driven by electric field, and in conductor *BDA* by nonelectric forces.





current performs work in section  $ACB$  at the expense of the work the nonelectric forces performed against the electric field.

Hence, in section  $BDA$  the electric energy is produced at the expense of other forms of energy and, conversely, in section  $ACB$  the electric energy turns into other forms of energy, for instance, into the internal energy of the conductor. Therefore the section of the circuit in which the charges move in the direction of the nonelectric forces is termed the *power source* (section  $BDA$ ), the term for the section in which the charges move in the direction of the electric forces being the *consumer of electric energy*.

A common power source in electrical engineering is the *generator*. Power sources that transform chemical energy into electric energy include *galvanic (voltaic) cells* and *storage batteries (accumulators)*.

The quantity characterizing the dependence of the energy gained by a unit charge in a generator on the properties and operating conditions of the generator is the *electromotive force* (emf) of the generator and the notation for it is  $\mathcal{E}$ . The measure for the electromotive force of a generator is the work of the nonelectric forces in transporting a unit positive charge  $q$

$$\mathcal{E} = W_{\text{nel}}/q \quad (18.8)$$

(Demonstrate that the unit for measuring emf in the SI system is 1 V.)

Thus, a continuous current can flow only in a circuit containing an emf. If the conductor  $ACB$  (Fig. 18.4) is withdrawn, the nonelectric forces will concentrate positive charges in  $A$  and negative charges in  $B$ . The voltage  $U$  between  $A$  and  $B$  will rise until the electric and the nonelectric forces become equal. Then the accumulation of charges in  $A$  and  $B$  will cease, the voltage between  $A$  and  $B$  reaching the maximum value for the given generator. (Explain why the voltage across an open-circuit generator is equal to its emf.)

Hence, to measure the emf of a generator it should be disconnected from the circuit and a voltmeter should be connected to it.

### 18-6 External and Internal Sections of a Circuit

It was demonstrated in the preceding section that an electric circuit consists of two essentially different sections. The term for the section of the circuit in which the charges

move in the direction of the electric forces ( $ACB$  in Fig. 18.4) is the *external section*, the term for the section in which the charges move in the direction of the nonelectric forces ( $BDA$  in Fig. 18.4) being *internal section*. In other words, the term internal section applies to the power source and the term external section to the rest of the circuit.

The boundary points between the internal and external sections are termed *poles*, or terminals ( $A$  and  $B$  in Fig. 18.4). The charges in the external section move from point to point only if there is a potential difference. Therefore, with the current flowing in the conductor, the potential in it decreases from point to point (in the direction from  $A$  to  $B$  in Fig. 18.4). Accordingly, the highest potential is that of one terminal and the lowest is that of the other (as compared with other points in the circuit). The terminal with the highest potential is termed *positive* and denoted by the sign  $+$ , while the terminal with the lowest potential is termed *negative* and denoted by the sign  $-$ .

In circuit diagrams use is made of schematic symbols shown in Figs. 18.3 and 18.5. A power source is denoted by two vertical parallel lines, the short thick line indicating the negative pole and the longer thin line the positive.

A schematic diagram of a simple electric circuit to which measuring instruments have been connected is shown in Fig. 18.6. We recall that the accepted direction of the current in the external circuit is that of the flux of positive charges from the positive to the negative pole (see Section 18-2), and in the internal circuit from the negative to the positive pole, despite the fact that the electrons in metals move in the opposite direction.

Since the potential in the external circuit decreases from point to point in the direction of the current, the voltage  $U$  across any section of the external circuit (see Fig. 18.6) will be less than the voltage  $U_{\text{ext}}$  at the power source terminals, that is, less than across the whole external circuit. Note that this is true only for a closed circuit. If the circuit is open, the potential of all the points of a conductor connected to one of the poles will be the same. (Ask yourself whether there will still be a voltage across the terminals.)

### 18-7 Ohm's Law for a Section of a Circuit Without EMF

In a section of a circuit carrying current there is a definite functional dependence between the current and the voltage across the section. It is termed the *current-voltage*, or *volt-ampere*, characteristic.

Fig. 18.5 Schematic symbols used in electric circuit diagrams.

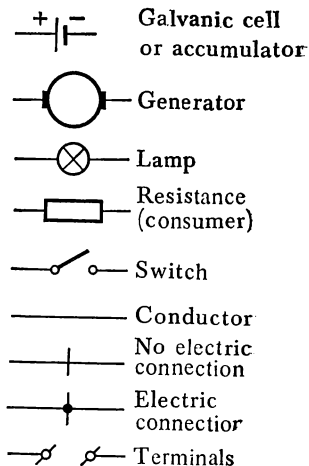
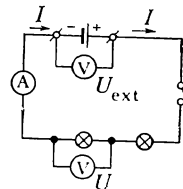


Fig. 18.6 Circuit diagram with ammeter and voltmeter.



The expression for the volt-ampere characteristic of a metallic conductor is (18.7b):

$$I = \frac{un_0eA}{l}U$$

It follows that there is a direct proportionality between  $I$  and  $U$ . This dependence was first established in experiments of the German scientist Georg Simon Ohm (1787-1854).

The volt-ampere characteristic is depicted as a  $I$  versus  $U$  diagram, which for a conductor, following (18.7b), takes the form of a straight line (Fig. 18.7). This dependence can be expressed by the formula

$$I = gU \quad (18.9)$$

where

$$g = \frac{un_0eA}{l} \quad (18.10)$$

The quantity  $g$  expressing the dependence of the current in a section of a circuit on its material and dimensions and on the surrounding medium is termed the *electric conductance* of the section. The measure for conductance is the current passing through the section when the voltage across it is unity.

In practice the relationship (18.9) is often written in the form

$$I = U/R \quad (18.11)$$

where

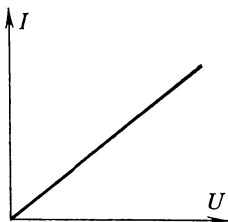
$$g = 1/R \quad (18.12)$$

The term for the quantity  $R$  is *electric resistance*. In the same way as friction in mechanics obstructs the motion of bodies the resistance of a conductor obstructs the directional motion of charges and causes the electric energy to be transformed into the internal energy of the conductor.

And so, the quantity characterizing the opposition to the current flowing in a conductor resulting from its internal structure and from the random motion of its particles is the electric resistance of the conductor. The measure of the resistance of a section of a circuit (without an emf) is the voltage across this section required to obtain a unit of current:

$$R = U/I \quad (18.11a)$$

Fig. 18.7 Volt-ampere characteristic of conductor.



We deduce a unit for measuring  $R$ :

$$R = \frac{1 \text{ V}}{1 \text{ A}} = 1 \frac{\text{V}}{\text{A}} = 1 \frac{\text{kg} \cdot \text{m}^2}{\text{s}^3 \cdot \text{A}^2} \\ = 1 \Omega \text{ (ohm, the Greek letter } \omega \text{)}$$

The unit of resistance in the SI system is the *ohm*. One ohm is the resistance of a section of a circuit in which a voltage of 1 V causes a current of 1 A to flow.\*

The regularity established by Ohm for metallic conductors is expressed by formula (18.11) and termed *Ohm's law for a section of a circuit without emf*: the current in a section of a circuit without emf is directly proportional to the voltage across the section and inversely proportional to its resistance,

$$I = U/R$$

An interesting conclusions can be drawn from this if the formula for Ohm's law is rewritten in the form

$$U = IR \quad (18.11b)$$

The physical meaning of this expression is that  $U$  is the total work performed by the electric field in moving a unit charge across this section of the circuit. In a section of a circuit of resistance  $R$  without emf all this work is spent on heating the conductor, that is, the energy is transformed into the internal energy of the conductor. It should be stressed once again that this transformation of energy is due to resistance, which acts in the same fashion as friction in mechanical processes.

If one assumes that the energy lost in the section is equal to the energy gained, one can regard (18.11b) as an expression of the law of energy conservation for this section. Accordingly, it can be said that the product  $IR$ , termed *voltage drop*, expresses the increase in the internal energy of the section, that is, it is numerically equal to the electric energy transformed into heat in the course of the passage of a unit charge across the section.

It follows from the above that if the electric energy in some section of a circuit is transformed into some other form of energy besides internal energy the voltage drop will be only a part of the voltage, that is, that the relation (18.11b) will not hold for such a section. In this case there will always be nonelectric forces acting in the section (there will be an emf).

\* The unit of conductance  $g = 1/R$  in the SI system is the *siemens* (S):  $1 \text{ S} = 1 \Omega^{-1} = 1 \text{ mho}$  (reciprocal ohm);

There is always a voltage drop in leads carrying current from a generator to a consumer. This is the reason why the voltage across the consumer is always less than across the generator's terminals.

### 18-8 Dependence of Resistance on Conductor's Material, Length and Cross Section

Let us now turn to factors that determine the resistance of a metallic conductor. The mobile charge carriers in a metal are the free electrons. The behaviour of electrons moving at random can be considered to be similar to that of gas molecules. Accordingly, in classical physics the term for the collection of free electrons is *electron gas*, it being assumed that in the first approximation it obeys the ideal-gas law.

The structure of the crystal lattice (and with it the density of the electron gas) depends on the metal. Therefore the resistance of a conductor must depend on its material. It must in addition depend on its length, cross section and temperature.

The cause for the dependence of the resistance on the cross section of a conductor is that a decrease in cross section brings about an increase in the density of the electron flux corresponding to the same current, with the result that the interaction between the electrons and the particles of the conductor is intensified. We can show this directly.

Since  $R = 1/g$  and  $g = un_0eA/l$  it follows that

$$R = \frac{l}{un_0eA}, \quad \text{or} \quad R = \frac{1}{un_0e} \frac{l}{A} \quad (18.13)$$

Denote

$$\rho = \frac{1}{un_0e} \quad (18.14)$$

Then

$$R = \rho l/A \quad (18.15)$$

It follows from this formula that the resistance of a conductor is directly proportional to its length and inversely proportional to its cross section. The term for the quantity  $\rho$  characterizing the dependence of the conductor's resistance on its material and on the surrounding medium is *specific resistance*, or *resistivity*, of the material. (Demonstrate that the unit of resistivity  $\rho$  in the SI system is  $1 \Omega \cdot \text{m}$ .) When doing calculations the resistivity of various materials is obtained from tables.

The quantity reciprocal to resistivity is termed *specific conductance*, or *conductivity*, of a substance and denoted  $\sigma$ :

$$\sigma = 1/\rho \quad (18.16)$$

(Demonstrate that the unit of conductivity  $\sigma$  in the SI system is  $1 \Omega^{-1}\text{m}^{-1} = 1 \text{ S/m.}$ )

### 18-9 The Temperature Dependence of Resistance

Since heating a conductor intensifies the random motion of its particles, it also increases the opposition to the directional motion of the charge carriers. This follows from formula (18.14):

$$\rho = \frac{1}{un_0e}$$

Heating a conductor results in a decrease in the charge carrier mobility  $u$  with  $n_0$  and  $e$  remaining constant, and  $\rho$  should increase accordingly. It has been established by experiment that within a wide temperature range the increment in a metal's resistivity is proportional to the increment in temperature. If we denote the resistivity at  $0^\circ\text{C}$  by  $\rho_0$  and at the temperature  $t$  by  $\rho_t$  we obtain  $\rho_t - \rho_0 = \alpha(t - 0)\rho_0$  or

$$\rho_t - \rho_0 = \alpha t \rho_0 \quad (18.17)$$

The quantity  $\alpha$  characterizing the dependence of the variation in resistivity of a heated conductor on its material is called the *temperature coefficient of resistance*. The measure for the temperature coefficient of resistance is the ratio of the variation of resistivity per  $1^\circ\text{C}$  to its value at  $0^\circ\text{C}$ :

$$\alpha = \frac{\rho_t - \rho_0}{t\rho_0} = \frac{\Delta\rho}{t\rho_0}$$

(Demonstrate that the unit used to measure  $\alpha$  is  $^\circ\text{C}^{-1}$ .)

The coefficient  $\alpha$  is positive for all metals since their resistance increases with temperature. The values of the temperature coefficients of resistance for pure metals are quite close and may all be assumed to be approximately equal to  $0.004^\circ\text{C}^{-1}$  (about  $1/273^\circ\text{C}^{-1}$ ). The resistivity of metal alloys is much greater than that of pure metals and the temperature coefficients of resistance are much less. This is because structural irregularities in their lattices due to the nonuniform distribution of atoms of component metals, which is independent of temperature, offer greater resistance to the motion of electrons than irregularities due

to the thermal motion of the lattice atoms. There are many alloys, such as constantan and manganin, whose  $\alpha$ 's are so small that their resistance can be regarded as independent of temperature.

Let us deduce a formula for calculating the resistance of conductors at various temperatures. Formula (18.17) yields

$$\rho_t = \rho_0 (1 + \alpha t) \quad (18.17a)$$

Substituting this value of  $\rho_t$  into (18.15), we finally obtain

$$R_t = \frac{\rho_0 l}{A} (1 + \alpha t) = R_0 (1 + \alpha t) \quad (18.18)$$

The temperature dependence of the resistivity of metals is utilized in *resistance thermometers*. They enable temperature measurements to be made with an accuracy of up to several thousandths of a degree (because resistance measurements can be very accurate).

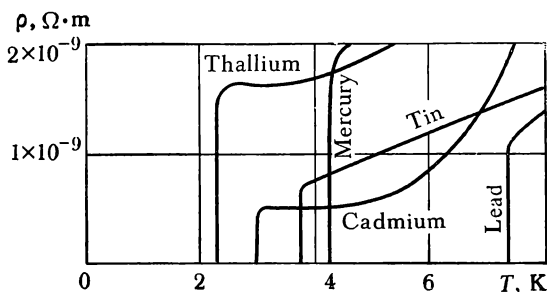
Note in addition that the coefficient  $\alpha$  for coal, electrolytes and pure semiconductors is negative because their resistance decreases with temperature (see Sections 24-1 and 24-2).

## 18-10 Superconductivity

The dependence of resistivity on temperature, it has turned out, cannot always be expressed with the aid of formula (18.18). Some interesting deviations from this relationship have been discovered in the extreme low temperature range. As the temperature of some conductors made of pure metals approaches absolute zero, their resistance tends not to zero, as predicted by (18.18), but some non-zero limiting value.

In 1911 the Dutch physicist Heike Kamerling Onnes (1853-1926), while measuring resistance in the extreme

**Fig. 18.8** Abrupt change in specific resistance in transition to superconducting state.



low temperature range, discovered the phenomenon later termed *superconductivity*. He established that in some cases the resistance of a substance at a definite temperature suddenly drops to zero (Fig. 18.8). If a closed circuit (for instance a ring) is made of such a substance and a current is induced in it, it will circulate for an indefinite time, for the charge carriers will not spend their energy on heating the conductor. The temperatures of transition to the superconducting state for several metals are presented in Table 18.1.

To obtain superconductivity one must have a substance with a regular lattice. Lattice imperfections make superconductivity impossible and such conductors have a finite resistance even at temperatures close to absolute zero.

Superconductivity makes it possible to pass enormous currents through conductors of small cross section at low temperatures. To this end the windings of powerful electric generators and high-power electromagnets are made of superconducting materials (niobium-titanium and niobium-tin alloys) cooled by liquid helium to 4 K. Superconducting cables to transport electric power are being developed.

**Table 18.1** Transition temperatures for some metals

Metal	$T$ , K
Lead	7.2
Tantalum	4.4
Mercury	4.22
Tin	3.71
Aluminium	1.14
Zinc	0.78
Magnesium	0.70

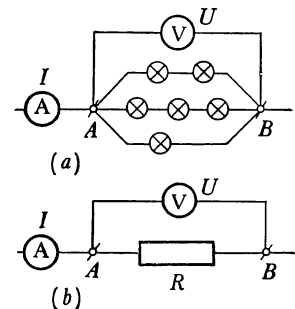
### 18-11 Equivalent Resistance

Suppose several consumers are connected in an arbitrary fashion between points  $A$  and  $B$  (Fig. 18.9a). Let the total current in these consumers be  $I$  and the voltage across points  $A$  and  $B$  be  $U$ . Fig. 18.9b shows the connection of a single conductor  $R$  between the points  $A$  and  $B$ .

Suppose the resistance  $R$  was chosen so as to make the readings of the ammeter and the voltmeter in diagrams (a) and (b) coincide. This means that if we substitute a single resistance  $R$  shown in Fig. 18.9b for the whole section of the circuit between points  $A$  and  $B$ , this will not result in any changes in the rest of the closed circuit (not shown in the figure). This means that the resistances between points  $A$  and  $B$  in both diagrams are equivalent.

The term for a resistance which when connected between two points of a circuit instead of all the other conductors leaves the current and the voltage unchanged is *equivalent resistance* of those conductors.

**Fig. 18.9** (a) Connection of consumers into circuit; (b) equivalent circuit diagram ( $R$  is equivalent resistance).



### 18-12 Electric Power Consumers in Series

An in-series connection of consumers is shown in Fig. 18.10. The currents, voltages and resistances for such a connection are calculated with the aid of rules presented below.

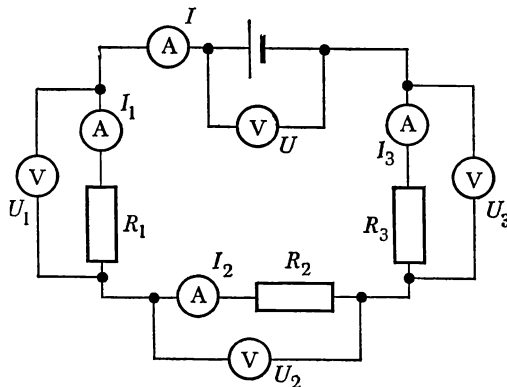


*Rule 1.* For a connection in series the current in all the sections of the circuit is the same:

$$I = I_1 = I_2 = I_3 \quad (18.19)$$

All the ammeters in Fig. 18.10 show the same current. The explanation is that the charges are neither generated nor destroyed in the circuit. In cases of an in series connection there is no sense in labeling the currents.

**Fig. 18.10** Connection of consumers in series.



*Rule 2.* For a connection in series the voltage across the external circuit is equal to the sum of the voltages across the individual sections of the circuit:

$$U_{\text{series}} = U_1 + U_2 + U_3 \quad (18.20)$$

This may be ascertained with the aid of voltmeters. (Explain relation (18.20) on the basis of the law of energy conservation.)

*Rule 3.* For a connection in series the voltages across individual sections of a circuit are directly proportional to their resistances:

$$U_1 \div U_2 \div U_3 = R_1 \div R_2 \div R_3 \quad (18.21)$$

(Prove this by making use of Ohm's law and relation (18.19).)

*Rule 4.* For a connection in series the equivalent resistance of the entire circuit is equal to the sum of the resistances of its individual sections:

$$R_{\text{series}} = R_1 + R_2 + R_3 \quad (18.22)$$

(Deduce this relationship using Ohm's law and formula (18.20).)

It follows from relation (18.20) that in the case of a connection in series of  $n$  identical circuit sections the total

voltage is

$$U_{\text{series}} = U_1 n \quad (18.23)$$

where  $U_1$  is the voltage across one section. In the same way we obtain from (18.22)

$$R_{\text{series}} = R_1 n \quad (18.24)$$

Note that when the circuit is disconnected at any one of the consumers connected in series, the current stops flowing in the circuit as a whole. This is the reason why this type of connection is not always convenient in practice.

### 18-13 Electric Power Consumers in Parallel

The parallel connection of consumers is shown in Fig. 18.11. Note that a point common to more than two conductors is called a *junction* (see Fig. 18.11). The term for all the conductors connected in parallel is *divided circuit*, the term for

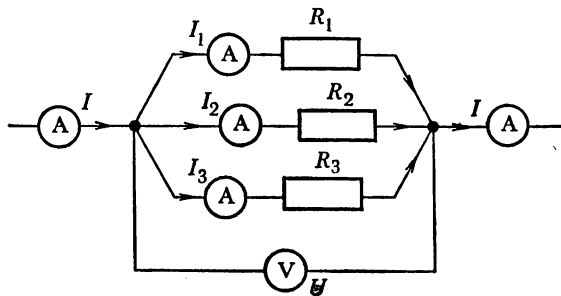


Fig. 18.11 Connection of consumers in parallel,

each being *branch*. There are also four rules used in calculating currents, voltages and resistances in the case of a connection in parallel.

**Rule 1.** For a connection in parallel the voltages across the individual branches and across the divided circuit as a whole are equal:

$$U_1 = U_2 = U_3 = U \quad (18.25)$$

**Rule 2.** The current in sections preceding and following the divided circuit is the sum of the currents in its individual branches:

$$I_{\text{par}} = I_1 + I_2 + I_3 \quad (18.26)$$

(Explain rule (18.26) making use of the law of charge conservation.)

**Rule 3.** The currents in the individual branches of a divided circuit are inversely proportional to their resistances:

$$I_1 \div I_2 \div I_3 = \frac{1}{R_1} \div \frac{1}{R_2} \div \frac{1}{R_3} \quad (18.27)$$

(Deduce this formula making use of Ohm's law and of formula (18.25).)

**Rule 4.** The conductance of a divided circuit is the sum of the conductances of its individual branches:

$$g_{\text{par}} = g_1 + g_2 + g_3 \quad (18.28)$$

or

$$\frac{1}{R_{\text{par}}} = \frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_3} \quad (18.28a)$$

(Deduce this relation using Ohm's law and formula (18.26).)

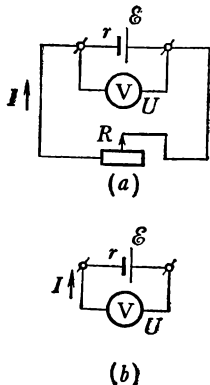
Note that the equivalent resistance of a divided circuit is always less than the smallest resistance of its branches. In the case when all the branches of a divided circuit are identical, the total current is

$$I_{\text{par}} = I_1 m \quad (18.29)$$

where  $I_1$  is the current in one branch and  $m$  is the number of branches. The equivalent resistance of a divided circuit is

$$R_{\text{par}} = \frac{R_1}{m} \quad (18.30)$$

**Fig. 18.12** (a) Increase in resistance of external circuit  $R$  results in increase in voltage  $U$  across the terminals of power source; (b) when emf is measured, voltmeter is external circuit.



If the voltage between the junctions remains constant, the currents in the branches will be independent of each other. For this reason parallel connection is convenient for most consumers.

### 18-14 Ohm's Law for a Complete Circuit

Suppose that an external circuit carrying a current  $I$  is connected to a power source with an electromotive force  $\mathcal{E}$  and that the voltmeter connected to the power source's terminals indicates the voltage  $U$  across the external circuit (Fig. 18.12a).

We recall that the power source is a conductor and because of this the current flowing in it generates heat. This generation of heat is due to the presence of a resistance  $r$  in the power source termed *internal resistance*. Applying the law of energy conservation, one draws the following conclusion.

The electromotive force  $\mathcal{E}$  is numerically equal to the energy gained by a unit charge in the internal circuit, the voltage  $U$  being equal to the energy it loses in the external circuit. Besides this, the charge loses the energy  $Ir$  in the

internal circuit, this being spent to generate heat in the power source.

Since the energy is neither generated nor destructed in the circuit, the charge gains as much energy as it loses in its journey along the entire circuit. Therefore

$$\mathcal{E} = U + Ir \quad (18.31)$$

If the external circuit is made up of static metallic conductors whose equivalent resistance is  $I$ , then  $U = IR$ , since in this case all energy is spent on heat. Substituting in (18.31)  $IR$  for  $U$ , we obtain

$$\mathcal{E} = IR + Ir \quad (18.32)$$

whence

$$I = \mathcal{E} / (R + r) \quad (18.32a)$$

This relation is called *Ohm's law for a complete circuit*: the current in an electric circuit containing an emf is directly proportional to the electromotive force and inversely proportional to the sum of the resistances of the external and internal circuits.

For a definite power source,  $\mathcal{E}$  and  $r$  in relation (18.31) can be regarded as constant. The external circuits connected to the power source may have different resistances  $R$ . Accordingly, the current  $I$  and voltage  $U$  will also be different. Since the sum  $U + Ir$  remains in this case constant, an increase in  $I$  will bring about a decrease in  $U$ , and vice versa.

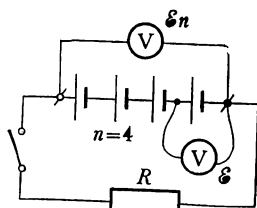
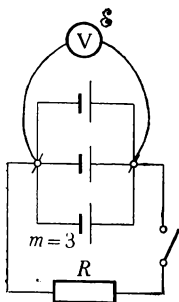
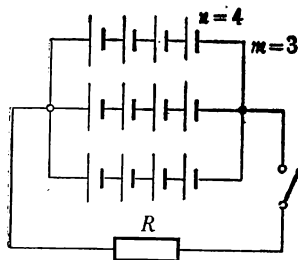
When  $R$  is very great as compared with  $r$ , the voltage drop in the internal circuit is so small as compared with  $U$  that it can be neglected. Hence, for great  $R$ 's the voltage across the external circuit  $U$  is approximately equal to the emf:

$$U \approx \mathcal{E} \quad (18.33)$$

This is the principle of measuring emf described in Section 18-5. Indeed, in the absence of an external circuit the power source is connected to a voltmeter (Fig. 18.12*b*), which indicates the voltage drop  $IR$  across itself equal to  $U$ . Since the resistance of a voltmeter is very great (see Section 25-17), relation (18.33) is valid in this case.

### 18-15 Combinations of Cells

Galvanic cells and accumulators used as power sources often have to be connected into batteries. They are connected in parallel, in series and in a combination of the two.

**Fig. 18.13** Connection of cells in series.**Fig. 18.14** Connection of cells in parallel.**Fig. 18.15** Combined connection of cells.

When cells are connected into a battery in series, the positive pole of the preceding cell is connected to the negative pole of the following cell (Fig. 18.13). Applying formula (18.32a) to the battery, one should keep in mind that in this case  $\mathcal{E}$  implies the emf of the battery as a whole,  $\mathcal{E}_{\text{bat}}$ , and  $r$  implies the internal resistance of the battery,  $r_{\text{bat}}$ . Hence, formula (18.32a) assumes the form

$$I = \frac{\mathcal{E}_{\text{bat}}}{R + r_{\text{bat}}} \quad (18.32b)$$

In practice the cells connected into a battery are always identical, since the use of dissimilar cells increases power losses and may be the cause of breakdown.

When  $n$  cells are connected in series, a charge acted upon by nonelectric forces gains energy in all the  $n$  cells in succession. Accordingly, the emf of the battery,  $\mathcal{E}_{\text{bat}}$ , will be equal to  $\mathcal{E}n$ , where  $\mathcal{E}$  is the emf of one cell. If the internal resistance of a cell is  $r$ , the internal resistance of the battery will be  $rn$ , since the charge overcomes the resistance of all the cells in turn. Hence, in the case of a connection in series of the cells Ohm's law for a complete circuit takes the form

$$I = \frac{\mathcal{E}n}{R + rn} \quad (18.34)$$

It follows from this formula that the connection of the cells in series results in substantial increase in the current only when the internal resistance of each cell is much less than the resistance of the external circuit,  $R$ .

For a parallel connection of the cells into a battery all the positive poles are connected to one terminal and all their negative poles to the other (Fig. 18.14). In this case the charge which passes through one cell does not pass through the others, that is, the emf of the battery,  $\mathcal{E}_{\text{bat}}$ , is equal to the emf of a single cell,  $\mathcal{E}$ , while the internal resistance  $r_{\text{bat}}$  of a battery made up of  $m$  identical cells is  $r/m$ . Hence, Ohm's law for a complete circuit in the case of a connection in parallel takes the form

$$I = \frac{\mathcal{E}}{R + r/m} \quad (18.35)$$

(Why is it expedient to connect the cells in parallel when the internal resistance of a cell is much greater than the resistance of the external circuit?)

The combined connection of cells is shown in Fig. 18.15. In this case only the connection in series increases the emf,  $\mathcal{E}_{\text{bat}} = \mathcal{E}n$ . If we take into account that connection in series increases internal resistance and that connection in

parallel reduces it, we obtain  $r_{\text{bat}} = (rn/m)$ . Hence, Ohm's law for the combined connection takes the form

$$I = \frac{\mathcal{E}n}{R + (rn/m)} \quad (18.36)$$

Analysis shows that the combined connection can be used to advantage when the resistance of the external circuit is close to the internal resistance of one cell.

### 18-16 Ohm's Law in General Form

It follows from experiments that nonelectric forces can act simultaneously in many sections of a closed circuit, including the consumer. This means that the generator is not the only possible location of emf's, for they are present in all sections of the circuit where nonelectric forces are active.

The emf of a section is assumed to be positive if the motion of the charges in this section coincides with the direction of the nonelectric forces. In this section a transformation into electric energy of other forms of energy must, of necessity, take place. If, however, the charges move in the direction opposite to that of the nonelectric forces, the emf is assumed to be negative. In this case the electric charges lose energy in overcoming the resistance of the nonelectric forces. The emf in such a section is termed *back emf*.

We recall that in overcoming resistance the charges lose energy, which is transformed into the internal energy of the conductor, that is, is spent on heating it. In the same way, in overcoming the opposition of nonelectric forces in the section of the circuit containing a negative emf, the charges lose energy, which is transformed into forms of energy other than internal energy. Hence, the presence of an emf in a section of the circuit results either in the transformation of other forms of energy into electric energy (in a generator) or in the transformation of electric energy into other forms of energy (except internal energy) in the case of a negative emf. For instance, in an electric motor the presence of a back emf makes the transformation of electric energy into mechanical possible.

To sum up, three different cases are possible when nonelectric forces act in a section of a circuit:

(1) the electric and the nonelectric forces act on the charges in opposite directions, the emf exceeding the voltage,

(2) the electric and the nonelectric forces act on the charges in opposite directions, the voltage exceeding the emf.

(3) the electric and the nonelectric forces act on the charges in the same direction.

Each of the above cases has its own peculiarities. Let us examine them.

The first is the case of a generator, that is, of the section of a circuit that supplies other parts of the circuit with electric energy. Its  $\mathcal{E}$  is numerically equal to the electric energy gained by a unit charge. If  $R$  is the resistance of the section as a whole (including the internal resistance of the generator  $r$ ),  $U$  the voltage across it and  $I$  the current in it, relation (18.31) should hold:

$$U = \mathcal{E} - IR$$

Indeed, it follows from the law of energy conservation that, if a unit charge has gained energy  $\mathcal{E}$  and spent energy  $IR$  on heating the conductor, its remaining energy is  $\mathcal{E} - IR$ . Finding  $I$  from (18.31), we obtain Ohm's law for a section of a circuit functioning as a generator:

$$I = (\mathcal{E} - U)/R \quad (18.37)$$

Note that the formula applies to any circuit connected to a generator.

The second case is that of an accumulator in the process of charging or of an electric motor. The electric energy spent by a unit charge in such a section is determined by the voltage,  $U = \varphi_1 - \varphi_2$  (Fig. 18.16). If the total resistance of the section is  $R$  and the current in it is  $I$ , the energy of a unit charge spent on heating the conductor will be  $IR$ , and the energy transformed into other forms of energy will be  $\mathcal{E}$ . Hence, in accordance with the law of energy conservation we have

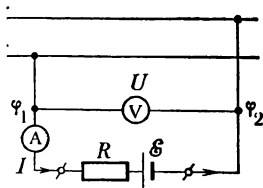
$$U = IR + \mathcal{E} \quad (18.38)$$

from which follows Ohm's law for a section of a circuit containing a back emf:

$$I = (U - \mathcal{E})/R \quad (18.38a)$$

In the third case the electric and the nonelectric forces act in the same direction. Accordingly, the charges move in the same direction. This means that such a section must,

**Fig. 18.16** Section of a circuit with an emf source connected into it;  $I < U/R$ .



of necessity, be a consumer of the energy gained in the rest of the circuit. It will, moreover, gain additional energy from the action of nonelectric forces. Hence, a unit charge having gained an energy  $U$  in other parts of the circuit gains additional energy  $\mathcal{E}$  in such a section and spends all its energy  $U + \mathcal{E}$  on heat, the latter energy being equal to  $IR$ . Therefore

$$IR = \mathcal{E} + U \quad (18.39)$$

Having determined  $I$ , we obtain Ohm's law for such a section of a circuit:

$$I = (\mathcal{E} + U)/R \quad (18.39a)$$

Combining all three cases, we can formulate Ohm's law for a section of a circuit containing emf as follows: the current in a section of a circuit containing emf is directly proportional to the algebraic sum of voltage and emf in this section and inversely proportional to its resistance.

In calculations where the sign of  $U$  or  $\mathcal{E}$  is unknown one should use formula (18.39). If the solution for  $U$  or  $\mathcal{E}$  prove to be a negative quantity, this means that their action on the charges is actually opposite to that assumed.

Note that for a connection in series of several emf's Ohm's law for a complete circuit assumes the form:

$$I = \sum_{i=1}^m \mathcal{E}_i / \sum_{i=1}^n R_i \quad (18.40)$$

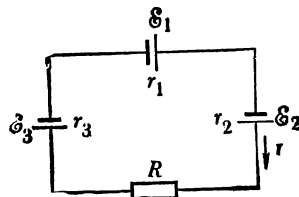
where  $m$  is the number of emf's, and  $n$  is the number of resistances (including the internal resistances of the emf sources) in the circuit.

An emf in a section of a circuit is positive if in this section the potential rises in the direction of the current. If the potential drops in this direction, the emf should be given a minus sign.

For the case of three emf's connected in series (as shown in Fig. 18.17) formula (18.40) must be written in the form:

$$I = \frac{\mathcal{E}_1 + \mathcal{E}_2 + \mathcal{E}_3}{r_1 + r_2 + r_3 + R}$$

Fig. 18.17 For in series connection of several sources, emf in closed circuit is equal to algebraic sum of all emf's





## 19

# Electric Power, Work and Heat Loss

## 19-1 Electric Current and Work

Let us see how the work of an electric current in a circuit is calculated.

The *total work of a current* in a consumer section of a circuit can be found from formula (17.10):

$$W_{\text{total}} = Uq$$

where  $U$  is the voltage across the section, and  $q$  is the charge transported over a cross section of the conductor in the time  $t$  the current was flowing in it. Since  $q = It$ , it follows

$$W_{\text{total}} = IUt \quad (19.1)$$

Since voltage and current in a section of a circuit can be measured with the aid of a voltmeter and ammeter, formula (19.1) can be conveniently used to compute the total work of the current, irrespective of the form of the energy into which the electric energy is converted in this section of the circuit.

For the case where all the electric energy is converted into internal energy (i.e. spent on heating a section of the circuit) formula (18.11) is valid:  $I = U/R$ . Substituting this expression into (19.1), we obtain another formula for computing the work of a current in a section without emf:

$$W = U^2 t / R \quad (19.2)$$

Since  $U = IR$ , formula (19.1) can be rewritten as

$$W = I^2 R t \quad (19.3)$$

Hence, for computing the work of a current in a section without emf any one of the formulae (19.1) to (19.3) can be used.

Consider now a section containing emf. Recall that when the consumer has a back emf, the electric energy is partly transformed into internal energy and partly into other forms of energy. The electric energy spent in this case is computed with the aid of formula (19.1). One question remains to be answered: Is it possible to compute what part of the electric energy turns into internal energy in such a section? The answer proves to be positive.

Since the  $IR$  drop is equal to the electric energy transformed into the internal energy per unit charge passing through the section, the increase in the internal energy of the section in the case of a charge  $q$  passing through it will be  $IRq$ , but since  $q = It$ , we obtain  $IRq = I^2Rt$ . Hence, the *work of a current spent on heat* in this section of the circuit is expressed by formula (19.3):

$$W_{\text{heat}} = I^2Rt$$

Note that this formula holds for all sections of a circuit, including the generator.

The work of nonelectric forces in the generator, used to estimate the amount of electric energy produced in it at the expense of other forms of energy, is found from expression (18.8). Since  $q = It$ , we have

$$W = \mathcal{E}It \quad (19.4)$$

Formula (19.4) can be applied to the consumer as well. In this case  $\mathcal{E}$  means a back emf, and the work  $W$  is the amount of electric energy transformed into mechanical or chemical energy.

We recall that, when making calculations in the SI system, the work obtained will be in joules (watt-seconds). In electrical engineering work is usually expressed in watt-hours or kilowatt-hours:

$$1 \text{ Wh} = 3.6 \times 10^3 \text{ J}, \quad 1 \text{ kWh} = 10^3 \text{ Wh} = 3.6 \times 10^6 \text{ J}$$

Since an hour consists of  $3.6 \times 10^3$  s, when calculating the work of a current in watt-hours, we only have to substitute into the formulae given above the time in hours, instead of seconds. Note that the name of the instrument measuring the work of a current is *wattmeter*, the term for the price of a unit of work being *tariff*. For instance, the tariff for the residents of Moscow is four kopecks per one kilowatt.

## 19-2 Power in a Direct Current Circuit

We recall that power is the term for a quantity characterizing the rate at which work is performed. The measure used for the power in a section of a circuit is the work performed by the current in this section per unit time. Since the usual notation for power in engineering is  $P$ , we have

$$P = W/t \quad (19.5)$$

(Demonstrate that the unit of power in the SI system is the watt:  $1 \text{ W} = 1 \text{ J/s}$ .)

Substituting the expressions for  $W$  from the formulae of the preceding section into (19.5), we obtain formulae for calculating power in electric circuits. The power of a current in a section without emf can be calculated with the aid of any of the following formulae (in calculations one should choose the formula that suits best):

$$P = UI \quad (19.6)$$

$$P = U^2/R \quad (19.7)$$

$$P = I^2R \quad (19.8)$$

When there is an emf,  $\mathcal{E}$ , in the consumer, the formula for total power is

$$P_{\text{total}} = UI \quad (19.6a)$$

while the formula for power spent on heat is

$$P_{\text{heat}} = I^2R \quad (19.8a)$$

The formula

$$P = \mathcal{E}I \quad (19.9)$$

determines the power spent on generating forms of energy others than internal. In the case of a generator, formula (19.9) determines the power spent in it generating electric energy.

In making calculations one should keep in mind that the power in the external circuit as a whole, irrespective of the type of connection is the sum of powers in individual sections of the circuit. Note that power spent on heating leads is often termed *heat loss*.

### 19-3 Heating Effects of Current

The first people to study the thermal effects of an electric current were the British physicist James P. Joule (1818-1889) and the Russian physicist Heinrich F. E. Lenz (1804-1865). The heat produced by an electric current flowing in a conductor is equal to the work of the electric field in overcoming the conductor's resistance:

$$= W_{\text{heat}} = I^2Rt \quad (19.10)$$

Formula (19.10) is the mathematical expression of *Joule's law*: the heat produced in a conductor by a current is directly proportional to the conductor's resistance, to the square of

the current and to the time it flowed in the conductor. Note once again that formula (19.10) makes it possible to calculate the heat produced by a current in any section  $R$  of the circuit.

In the case of a connection in series of conductors whose resistances are  $R_1$  and  $R_2$  (Fig. 19.1a) the heat produced in them can be expressed as follows:

$$Q_1 = I^2 R_1 t, \quad Q_2 = I^2 R_2 t \quad (19.11)$$

whence

$$Q_1/Q_2 = R_1/R_2 \quad (19.12)$$

Therefore in a series connection the heat produced in each conductor is directly proportional to its resistance.

The heat produced by a current in the case of a parallel connection of two sections with resistances  $R_1$  and  $R_2$  without emf's (Fig. 19.1b) can be expressed for each section separately as follows:

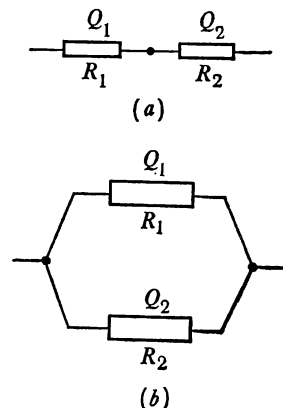
$$Q_1 = U^2 t/R_1, \quad Q_2 = U^2 t/R_2$$

whence

$$Q_1/Q_2 = R_2/R_1 \quad (19.13)$$

Thus, the heat produced by a current in sections without emf's and connected in parallel is inversely proportional to their resistances.

**Fig. 19.1** (a) Connection in series results in greater amount of heat being liberated in conductor with larger resistance; (b) connection in parallel results in greater amount of heat being liberated in conductor with least resistance.



#### 19-4 Relation of Resistance to Heating Effect

The connection of a conductor of very small resistance across a generator's terminals is termed a *short circuit*. In this case the current is limited only by the internal resistance of the generator,  $r$ .

Indeed, when  $R$  is much less than  $r$ , one can assume the resistance  $R$  of the conductor short-circuiting the generator to be zero. We then obtain from formula (18.32a),  $I = \mathcal{E}/(R + r)$ ,

$$I_{max} = \mathcal{E}/r \quad (19.14)$$

This formula gives the maximum current a power source with an emf and an internal resistance  $r$  can deliver. The term for  $I_{max}$  is *short-circuit current*.

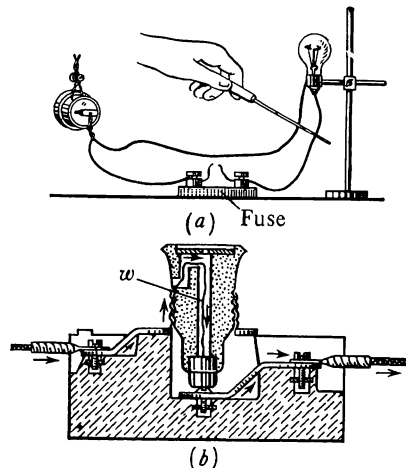
A short circuit is harmful. Apart from wasting energy, a short circuit can put a generator out of action or cause a fire as the result of the overheating of the wires short-circuiting the generator. Short circuit can be the result

not only of the direct contact of wires connected to the generator's terminals but also of indirect contact through low-resistance structural elements, the ground, etc. Therefore wiring should be well insulated from them.

The heating effects of current are widely used in technology and everyday life. One of the most common applications is for lighting rooms, shops, etc., with the aid of *incandescent filament lamps*. The first such lamp was made by A. N. Ladygin, while the American inventor Thomas A. Edison (1847-1931) made it an article of universal use.

The operation of such familiar devices as electric heaters, kettles, irons, soldering irons, stoves, etc., is based on the heating effect of an electric current. In cases where the heating effects are the cause of power losses efforts are made to reduce them. For instance, the liberation of heat in leads is harmful. To reduce its resistance of the leads and the current flowing in them are made as small as possible. Electric power is transmitted at high voltages, which makes it possible, while leaving the power unchanged, to reduce the current (in accordance with (19.6)) and the losses in heating the wires.

Fig. 19.2 (a) In short circuit wire strip in fuse melts and circuit is broken; (b) commercial electric fuse.



To prevent fires and damage to generators caused by short-circuiting, *fuses* are always connected into an electric circuit (Fig. 19.2a). Note that the resistance of a fuse wire per unit length should be much greater than that of the leads. This can be seen from expression (19.11).

Figure 19.2b illustrates the design of a fuse. The current supplied to the consumer (indicated by arrows) passes

through a strip of wire  $w$  made of low-melting alloy. In the case of short-circuiting, the current suddenly increases and the heat produced by it melts the strip of wire, thus breaking the circuit.

# Thermoelectricity

# 20

## 20-1 Thermionic Emission

All metals have free electrons which move at random between the positive ions making up the crystal lattice. Inside the metal the action of the positive ions on the electrons is, on the whole, compensated, but an electron outside the external layer of the positive ions is attracted by them. When electrons moving at random cross the surface of the metal, this force restricts their motion and draws them back into the metal. This means that the potential energy of an electron in a metal is less than outside it. Hence, if the potential energy of an electron outside a metal is assumed to be zero, its energy inside the metal will be negative.

Suppose we take a piece of metal, shown in Fig. 20.1a as a shaded rectangle. Choose an axis  $x$  perpendicular to the metal's surface with the origin at 0. The variation in the electron's potential energy with  $x$  is illustrated in Fig. 20.1b. The diagram has the shape of a potential well. Here  $\Delta P.E.$  is the jump in potential energy of a free electron leaving the metal, that is, the depth of the potential well which contains electrons moving at random inside the metal. To leave the metal the electron must perform work  $\Phi$  equal in magnitude to the depth of the potential well at the expense of its kinetic energy:

$$\Phi = \Delta P.E. \quad (20.1)$$

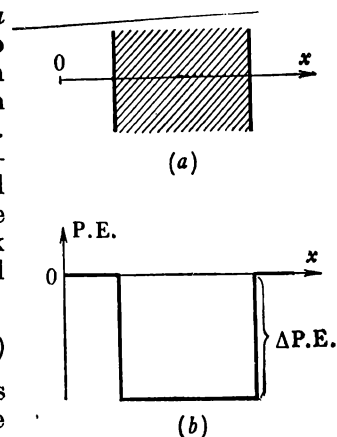
Since the jump in the potential energy of an electron is due to the electric field existing at the boundary of the metal, it follows that

$$\Delta P.E. = e\Delta\phi \quad (20.2)$$

where  $\Delta\phi$  is the potential jump across the metal's surface, and  $e$  is the electron charge. Substituting the expression for  $\Delta P.E.$  from (20.2) into (20.1), we obtain

$$\Phi = e\Delta\phi \quad (20.3)$$

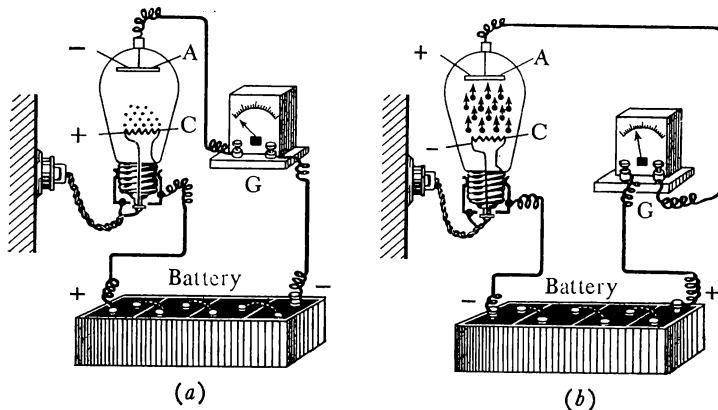
Fig. 20.1 Potential energy of electron in metal.



The minimum work  $\Phi$  that an electron must perform at the expense of its kinetic energy to leave a metal (and stay away) is called the *work function*. It can be assumed for all practical purposes that the work function depends only on the nature of the metal and on the condition of its surface.

Under normal conditions the average kinetic energy of the random motion of free electrons in a metal is much less than  $\Phi$ . However, some electrons whose kinetic energy exceeds  $\Phi$  may still be able to leave the metal. Therefore there are always electrons moving at random above the

Fig. 20.2 (a) Filament C emits electrons which, if anode A is charged negatively, keep close to filament; (b) if anode A is charged positively, electrons go to it, causing pointer of galvanometer G to deflect.



surface of a metal. Since the average kinetic energy of free electrons in a metal increases with temperature, one can expect the number of electrons emitted by its surface to become quite high at sufficiently high temperatures. Experiments proved this to be true. The process of electrons leaving a heated metal is known as *thermionic emission*. A perceptible emission of electrons from metals starts at temperatures of about 1000 K.

Take an incandescent lamp with an additional electrode A sealed into it (Fig. 20.2). Connect this electrode and the filament to a battery in series with a galvanometer G. If the filament is not connected to another power source, there is no current flowing through the galvanometer, since the cold filament C emits practically no electrons. When the lamp's filament, termed the *cathode*, is connected to its power source, the current flows through the galvanometer, but only if electrode A, termed *anode*, or *plate*, is connected to the positive terminal of the battery (Fig. 20.2b). There is no current if electrode A is connected to the negative

terminal (Fig. 20.2a). This means that the filament emits negative charges, electrons, which are the charge carriers in this case.

## 20-2 Contact Potential Difference

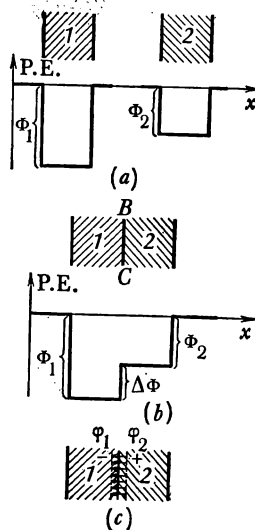
Let us examine the process of electrifying two different metals brought into contact with each other (see Section 16-5). There are two causes of such electrification. The first is the difference in the work functions of the metals, the second the difference in the densities of the electron gas in the metals, or the difference in the number of free electrons per unit volume in each metal. Let us discuss the effect of the difference in work functions.

Suppose we have plates 1 and 2 made of different metals such that  $\Phi_1 > \Phi_2$  (Fig. 20.3a). We bring the plates in contact. Here the diagram of the potential energy will be as shown in Fig. 20.3b. It can be seen that for an electron to go from metal 1 to metal 2 it must perform work  $\Delta\Phi$  much less than  $\Phi_1$ . Because of this many electrons will go from left to right even at room temperature. The situation for the electrons in metal 2 is such that all the electrons which in their thermal motion manage to cross the boundary BC will remain in metal 2, since their potential energy decreases in the process.

One can picture the whole process as follows: in going from metal 1 to metal 2 the electrons must overcome a potential barrier, while in going from metal 2 to metal 1 they "roll down" the potential barrier  $\Delta\Phi$  by themselves. This means that the electrons crossing from right to left should be more numerous than the electrons crossing in the opposite direction. Therefore metal 1 acquires a negative charge and metal 2 a positive, that is, an electric field is established between them. The entire field is concentrated in a thin transition layer separating the metals (Fig. 20.3c). This field prevents electrons from crossing from metal 2 to metal 1. The result is that the electron fluxes in both directions soon become equal and a dynamic equilibrium sets in between the metals.

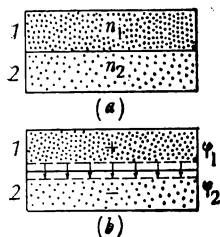
The potential difference  $\varphi_2 - \varphi_1 = \Delta\varphi$  established between the contacting metals as the result of the dynamic equilibrium of their electrons is called the *contact potential difference*. Contact potential difference—the result of the difference in the electron work functions of the contacting metals—can be as high as several volts, its value being practically independent of temperature.

Fig. 20.3 (a) Potential curves for metals 1 and 2 before equilibrium is established; (b) and (c) in equilibrium, transition of electrons from metal 1 to metal 2 generates electric field in transition layer.





**Fig. 20.4** (a) Density of electron gas is greater in metal 1 than in metal 2; (b) on contact prevailing diffusion of electrons from metal 1 to metal 2 causes contact potential difference to be established between metals.

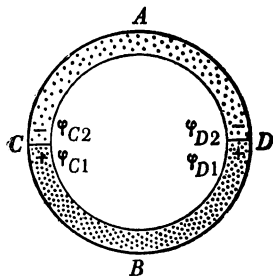


Let us now examine the part played by the second factor. Let the free electron concentration in the first metal,  $n_1$ , exceed that in the second:  $n_1 > n_2$  (Fig. 20.4a). To simplify the discussion we assume the work functions of both metals to be equal. In that case the electron flux from metal 1 to metal 2 due to electrons' random motion will exceed the flux in the opposite direction; metal 1 will acquire a positive charge and metal 2 a negative (Fig. 20.4b). The resulting electric field in the transition layer prevents the electrons from crossing from metal 1 to metal 2, and a state of dynamic equilibrium is soon established between the electron fluxes flowing in opposite directions.

The contact potential difference in metals,  $\Delta\phi$ , which is due to the difference in their free electron concentrations, does not exceed more than several hundredths of a volt and rises with the temperature of the contacting metals. The explanation is that heating makes the electrons move more quickly, and this brings about an increase in the electron fluxes flowing in both directions. The difference in both fluxes increases in proportion to the increase in each of them. Therefore the potential difference  $\Delta\phi$  between the metals, and hence the field in the transition layer, increase.

### 20-3 Thermoelectromotive Force

**Fig. 20.5** If different metals are at equal temperatures, contact potential differences at both junctions are equal in magnitude and opposite in sign and no current flows in circuit.



Contact potential difference cannot be the cause of a current flowing in a closed circuit all sections of which are at an equal temperature, because it merely compensates electron fluxes flowing in opposite directions (see Section 20-2). Taking the algebraic sum of all the potential variations at the contacts in such a circuit, we will find it to be zero. This means that under such conditions the contact potential difference does not constitute an emf (Fig. 20.5). But if the contacts  $C$  and  $D$  are at different temperatures, there will be an emf in the circuit.

Indeed, if the temperature of the contact  $D$  is raised, electrons from metal  $B$  will cross it to enter metal  $A$ , with the result that the contact potential difference at junction  $D$  rises. An increase in the number of electrons in metal  $A$  at point  $D$  will cause them to flow towards point  $C$ . An increase in their concentration at point  $C$  will, in turn, lead to their crossing contact  $C$  from metal  $A$  to metal  $B$ . From this point they will travel along metal  $B$  to reach contact  $D$ . Hence, if the temperature of contact  $D$  is kept higher than that of contact  $C$ , there will be a directional motion of

electrons in the anticlockwise direction. This means that there is an emf in the circuit.

An emf in a closed circuit made up of different metals resulting from the difference in the temperatures of the contacts is called *thermoelectromotive force*. Thermoelectromotive force in a circuit made up of two different metals is directly proportional to the difference between the temperatures of their contacts and depends on the nature of the metals. Electric energy in such a circuit is produced at the expense of the internal energy of the source maintaining the temperature difference.

Note that thermoelectromotive force is not great, its value for metals being no more than several hundredths of a thousandth of a volt per degree of temperature difference in the contacts in the circuit. Semiconductors have much greater thermoelectromotive force (up to a thousandth of a volt per degree). There are two reasons for this. The first is that the energy of charge carriers in semiconductors strongly depends on temperature and the second is that semiconductors with a considerable difference in their electron concentrations can be chosen as the pair.

A device made of two different metals fused together at one end and used to generate electric energy at the expense of internal energy of an external body which maintains

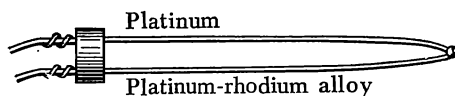


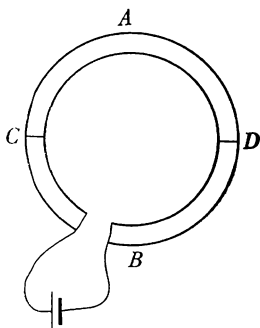
Fig. 20.6 Thermocouple.

the temperature difference between the junctions of the device is termed a *thermocouple*. In a thermocouple one junction is made by fusing together ends of wires (or plates) made of different metals (Fig. 20.6), the free ends being connected to an external circuit or to measuring instruments. The part of the second (cold) junction is played by the contacts of the thermocouple with the leads.

#### 20-4 The Peltier Effect

Let us see what happens if a power source is connected to a circuit made up of two different metals *A* and *B* (Fig. 20.5), the power source establishing a current coinciding in direction with that established by the heated contact *D* (Fig. 20.7). In this case the electron flux will be retarded at junction *D*, because the electrons have to overcome the contact

Fig. 20.7 If current flows in circuit made up of different metals, one junction is heated and other is cooled.



potential difference in the transition layer of contact *D*. The situation will be different in junction *C*, where the electrons will be accelerated, since the forces in the transition layer of this contact act on them in the direction of their flow. Hence, in contact *D* the kinetic energy of the electrons will be transformed into potential energy while in contact *C* their potential energy will be transformed into kinetic energy. This means that when the circuit shown in Fig. 20.7 is closed, the temperature of contact *D* must drop and that of contact *C* must rise. Experiments proved this to be true. A reversal of the current brings about the cooling of contact *C* and the heating of contact *D*. (Explain why.)

This phenomenon was discovered in 1834 by the French watchmaker and scientific amateur Jean C. A. Peltier (1785-1845) and bears his name. The *Peltier effect* in metals is very small and metallic thermocouples cannot be used to achieve substantial cooling. It is much more pronounced in semiconductors. This fact has made possible its practical application in thermoelectric coolers, distinguished by their simplicity of design and compactness.

The Peltier effect is used in medicine for cooling the instrument used for extracting the crystalline lens out of the eye. After the operation has been performed, the lens is replaced in the eye with the aid of the same instrument. The point of the instrument consists of a junction of two different semiconductors cooled by passing a current at such low temperatures that the crystalline lens, upon contact, freezes on to it. A reversal of the current frees the lens.

### 20-5 Application of Thermoelectricity in Science and Technology

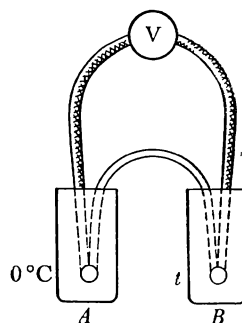
It was mentioned above that the emf of a thermocouple is directly proportional to the difference in temperature of junctions made of different metals. There are some deviations from this rule due to changes in the internal structure of some metals at certain temperatures. Still, thermocouples obeying this rule can be chosen for any desired temperature range. Hence, by measuring the emf of a thermocouple one can determine the temperature difference between its junctions, in other words, use the thermocouple to measure temperature. The accuracy of such measurements is determined by the accuracy of measuring emf's with a voltmeter.

The existence at present of accurate and sensitive voltmeters makes it possible to measure very small temperature

differences with the aid of thermocouples. Besides, thermocouples can be used to measure both high and very low temperatures and because of this are widely used in science and industry as precision thermometers. One such thermocouple is illustrated in Fig. 20.8. One of the junctions (*A*) is immersed into melting snow, and the temperature of the other (*B*) is measured with the aid of a voltmeter with a scale calibrated in degrees.

Thermocouples are used not only to measure temperature but also to control it, for they supply information about temperature in the form of an electric signal (thermoelectromotive force), which can be easily amplified and used to control the power of a heater. Sometimes to increase sensitivity several thermocouples are connected into a *thermopile* (Fig. 20.9). This enables measurements of very weak fluxes of radiation (for instance, coming from a star) to be made. High temperatures are measured with thermocouples made of refractory metals (for instance, of platinum and its alloys (see Fig. 20.6)).

Fig. 20.8 Measuring temperature with thermocouple.



## Electric Current in Electrolytes

## 21

### 21-1 Electrolytic Dissociation

We now consider currents passing through solutions of acids, salts and alkalis.

Pure distilled water is essentially a dielectric. This can be demonstrated in the following experiment: if an incandescent lamp is connected in series with a basin filled with distilled water into which metal plates have been immersed, and the circuit is then connected to a d.c. source, the lamp will not burn. A solution of sugar in water will not conduct current either. But if a few drops of acid from a pipette are added to the water in the basin, the lamp will light up brightly. Hence, a solution of acid in water is a good conductor of electricity. Let us discuss the causes of these phenomena.

It was shown in Section 17-10 that water molecules are natural dipoles. Suppose water contains a molecule of hydrochloric acid  $\text{HCl}$ . This molecule is made up of the ion  $\text{H}^+$  and the ion  $\text{Cl}^-$  held together by Coulomb attraction forces. We recall that water greatly weakens (by about 80 times) the interaction of charges and that the water molecules moving at random strike the molecules of hydrochloric

Fig. 20.9 Schematic diagram of thermopile.

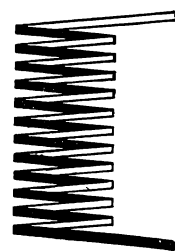
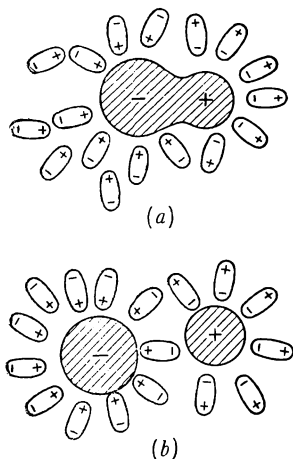
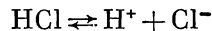


Fig. 21.1 Dissociation of HCl molecule in water.



acid from all sides; the result is that the HCl molecules dissociate into ions. The water dipoles may be said to surround the acid molecule and to pull their ions apart (Fig. 21.1). Note that ions of opposite sign in water are attracted and on meeting may again form a molecule. Therefore not only the process of dissociation of acid molecules takes place in water, but the reverse process of the formation of neutral molecules from ions as well:



(the arrows indicate that the process works in both directions). The disintegration of molecules into ions as a result of solution is called *electrolytic dissociation*. The ratio of the number of molecules that dissociated into ions to the total number of solute molecules is sometimes called the *dissociation ratio*.

It follows that only ions are mobile charge carriers in solutions. It is essential that the hydrogen and metal ions formed in the process of dissociation are always positive. Ions in a solution may sometimes consist of a group of several atoms.

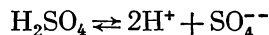
Note, in addition, that the presence of a solvent is not the only cause of the dissociation of a molecule into ions. This may, for instance, be the result of heating the substance to high temperatures. This is why fused salts also conduct electricity.

Hence, according to the theory of electrolytic dissociation there are always free ions in solutions of acids, salts and alkalis, because they are produced when these substances are dissolved in water or some other solvent.

## 21-2 Electrolysis

Let us examine in more detail the passage of current through a solution containing mobile ions.

A liquid conductor that contains no mobile charge carriers other than the ions is called an *electrolyte*. Let there be a solution of sulphuric acid in water in a basin. The dissociation of sulphuric acid molecules in water takes place in compliance with the equation



We immerse platinum plates into the solution and connect them to a battery (Fig. 21.2) through an ammeter. Such plates are termed *electrodes*. The electrode connected

to the positive terminal is termed the *anode* and the one connected to the negative the *cathode*. If the switch is turned on, an electric field will be established in the electrolyte between the electrodes. Acted upon by the forces of this field the hydrogen ions  $H^+$  start moving to the cathode and the ions of the acid radical  $SO_4^{--}$  to the anode. On reaching the cathode each  $H^+$  ion borrows a free electron from the electrode plate and turns into a neutral hydrogen atom. Two such atoms associate to form a molecule of gaseous hydrogen, which is liberated at the cathode.

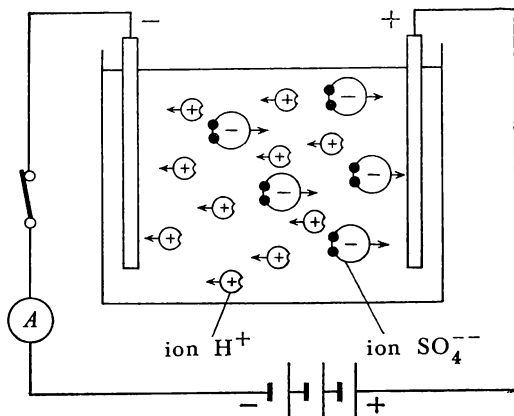


Fig. 21.2 Movement of ions in electrolytic solution.

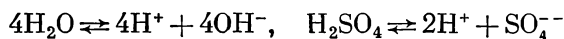
It turns out that in the case being discussed there are other negative ions in the electrolyte besides the  $SO_4^{--}$  ions, for the molecules of the water itself also dissociate in small numbers:

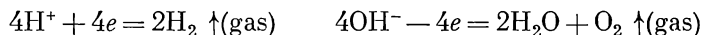


The  $OH^-$  (hydroxyl) ions easily give up their extra electron, the  $SO_4^{--}$  retaining theirs more firmly. Because of this the  $OH^-$  ions are discharged at the anode (at which the negative ions arrive), the  $SO_4^{--}$  ions remaining in the solution. The discharge of the  $OH^-$  ions results in the formation of water and neutral molecules of gaseous oxygen, which is liberated at the anode.

Denoting the magnitude of the electron charge by  $e$ , we can write all those processes in the following form:

*Formation of Ions in Solution*



*Cathode Reaction\***Anode Reaction*

It follows that the component parts of water leave the solution, while those of sulphuric acid remain in it. This means that as the current continues to flow, the amount of water in the solution decreases and the concentration of the solution rises. This is why this process is sometimes termed *decomposition of water by electric current*.

It follows from the above that the current flowing through an electrolyte is accompanied by transformations of substance, that is, the current in electrolytes is chemically active. The process of electric current passing through an electrolyte and accompanied by chemical transformations and by the depositing of substances at the electrodes is called *electrolysis*. A vessel containing an electrolyte and equipped with electrodes is termed an *electrolytic cell*.

The name for the positive ions in an electrolyte is *cations* (because in electrolysis they move to the cathode), while that for the negative ions is *anions*. Recall that hydrogen and metal ions are cations.

In the example given above substances are deposited at both electrodes. It will become clear in the following section that this is not always the case. The depositing of substances at both electrodes takes place only in the case of a passive anode, which does not itself dissolve in the electrolyte. The plates in our example do not react with the electrolyte. The electrodes most frequently used for such processes in industry are made of coal or graphite.

The current in an electrolyte obeys Ohm's law, that is, it varies in direct proportion to the voltage. The heating of electrolytes brings about a decrease in their viscosity and an increase in the mobility of the ions. Further, the dissociation ratio of the solute molecules rises with temperature, leading to an increase in the number of charge carriers contained in the electrolyte. This means that the resistance of electrolytes drops with an increase in temperature.

### 21-3 Electrolysis Involving Anode Dissolution

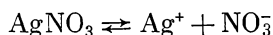
It was mentioned above that electrolysis is not always accompanied by the depositing of substances at both electrodes. If the electrolytic cell contains a solution of the

\* The vertical arrow means that gas is liberated from the solution.

salt of the substance of which the anode is made, deposits will appear only at the cathode, and the anode will dissolve. In this case the anode is said to be *active*, since it reacts with the electrolyte. Let us consider the process of extracting foreign impurities from silver, the process known as *refining*.

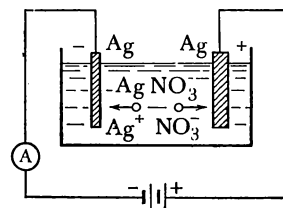
Silver electrodes are immersed into a cell containing a solution of silver nitrate and connected to a battery (Fig. 21.3). The cathode is a thin plate of pure silver, the anode a thick plate of silver containing impurities. In the process of electrolysis the silver is transported from the anode to the cathode and the impurities precipitate from the solution and sink to the bottom. Let us see how this happens.

The molecules of silver nitrate dissociate:



In the course of electrolysis each silver ion reaches the cathode and borrows one electron from it. Thus neutral silver atoms are deposited on the cathode so that its mass grows in the process. The  $\text{NO}_3^-$  ions move to the anode, where each joins a silver ion dissolved in the solvent from the anode. Hence, the mass of the anode gradually diminishes in the process of electrolysis, the concentration of the solution in the bath remaining constant until the anode is completely dissolved. (Explain why the concentration of the solution remains unchanged.)

Fig. 21.3 Electrolysis involving anode dissolution.



#### 21-4 Faraday's First Law

The phenomenon of electrolysis was first studied by Faraday. Measuring the charge transported through the electrolyte and the mass of the cathode before and after electrolysis, he established that the mass of the material deposited in the process of electrolysis is directly proportional to the amount of electricity that passed through the solution:

$$m = kq \quad (21.1)$$

Formula (21.1) is the mathematical expression of *Faraday's first law*.

Faraday's experiments demonstrated that the mass of the material deposited in the course of electrolysis is determined not only by the charge  $q$  but by the material as well. The proportionality factor  $k$  expressing the dependence of the mass of the material deposited in the course of electro-



lysis on its nature is called the *electrochemical equivalent* of the material. The measure of the electrochemical equivalent is the mass of the material deposited on one electrode per unit charge passing through the electrolyte:

$$k = m/q \quad (21.1a)$$

Electrochemical equivalents are usually expressed in gram per coulomb (g/C).

Since  $q = It$ , Faraday's first law can be written in the form

$$m = kIt \quad (21.2)$$

Electrochemical equivalents can be measured in experiment with great accuracy. At one time this fact was used (21.1) to define the coulomb in terms of the electrochemical equivalent of silver, which was measured with great accuracy and was found to be  $1.118 \times 10^{-3}$  g/C = 1.118 mg/C.

Let us see how theory explains the results of Faraday's experiments. An ion, while discharging in the process of electrolysis at the cathode, borrows a definite number of electrons (for instance, a silver ion borrows one electron, while a copper ion borrows two). Therefore the charge transported through the electrolyte must be proportional to the number of ions discharged. Since all masses of the ions of some material are the same, their total mass is proportional to their number. This means that the mass of a material deposited during electrolysis is directly proportional to the amount of electricity that has passed through the solution, as demanded by Faraday's first law.

### 21-5 Faraday's Second Law

One mole of ions discharged at an electrode is the mass of these ions in grams numerically equal to the relative molecular mass of one ion. The *chemical equivalent*, or *gram-equivalent*, of these ions is the term for the ratio of a mole of them to their valency. We denote the relative molecular mass of an ion by  $A$  and its valency by  $n$ . The chemical equivalent is  $A/n$ .

For example, the relative atomic mass of copper is 63.54 and the valency of copper ions is 2, a mole of copper makes 63.54 g and its chemical equivalent is  $(63.54/2) = 31.77$ g. (Find the chemical equivalent of the  $\text{NO}_3^-$  ions if the relative atomic mass of oxygen is 16 and that of nitrogen is 14.)

Faraday established as a result of his experiments that an identical amount of electricity  $F$  is required to deposit

at an electrode a chemical equivalent of ions of any kind. The term for this amount of electricity is the *faraday*. According to modern data  $F = 9.652 \times 10^4$  C. Therefore the total charge of all the ions in a chemical equivalent is  $F$ .

The ratio  $m \div (A/n)$  of the number of grams,  $m$  deposited at an electrode, to the chemical equivalent,  $A/n$ , is the number of chemical equivalents deposited in electrolysis. The ratio  $q \div F$  of the electron charge transported through the solution in coulombs,  $q$ , to the amount needed to deposit one chemical equivalent,  $F$ , is also the number of chemical equivalents deposited. Hence  $m \div (A/n) = q \div F$  and

$$m = \frac{A}{nF} q \quad (21.3)$$

Comparing (21.3) with (21.1), we obtain

$$k = \frac{1}{F} \frac{A}{n} \quad (21.4)$$

Formula (21.4) is the mathematical expression of *Faraday's second law*: the electrochemical equivalents of various substances are directly proportional to their chemical equivalents.

Note that formula (21.3) expresses the combined Faraday law for electrolysis.

Let us now see how the charge of a monovalent ion, or the electron charge  $e$ , can be determined with the aid of Faraday's laws. If the valency of the ions is unity ( $n = 1$ ), their chemical equivalent  $A/n$  will be equal to their relative molecular mass  $A$ , the charge of each ion being  $e$ . We recall that the number of ions in a mole is equal to the Avogadro number  $N_A$  (see Section 3-6). Therefore, if the total charge of the ions is  $F$ , it follows that

$$e = F/N_A \quad (21.5)$$

Substituting the numerical values of  $F$  and  $N_A$  into (21.5), we obtain the charge of a monovalent ion or electron:

$$e = \frac{9.652 \times 10^4 \text{ C/mol}}{6.02 \times 10^{23} \text{ ions/mol}} = 1.60 \times 10^{-19} \text{ C/ion}$$

This value of the electron charge is in good agreement with the results of Millikan's experiments and speaks for both the electron theory of the structure of matter and the theory of electrolytic dissociation.

### 21-6 Some Applications of Electrolysis

The uses of electrolysis in technology are manifold. Let us cite several examples.

It was mentioned above that electrolysis is used to purify metals smelted from ores. The electrolysis of melted ores is used to produce from them light metals, which react with water and cannot be precipitated out of aqueous solutions. This method is used in the production of aluminium, sodium, lithium, etc. Zinc and nickel are produced by the *electroextraction method*, that is, the metals are extracted from solutions with the aid of electrolysis. Atomic oxygen liberated in the process of electrolysis is a very powerful oxidizer. It is used in the production of various drugs (for instance, iodoform).

Electrolysis is used to deposit thin films of corrosion-resistant metals on objects made of other metals to protect them from atmospheric corrosion. This method of coating is termed *electroplating*. Nickel and chrome plating are relevant examples. Electroplating is used in making jewellery, for instance for silver and gold plating. The term used for the process of manufacturing relief copies of pictures with the aid of electrolysis is *galvanoplastics*. It was invented in 1837 by the Russian scientist B. S. Jakobi (1801-1874). This method is used in making printers' plates and reproducing matrices for printing books, newspapers, etc.

Since an electric field is greater at points and protrusions of a metallic surface of an electrode, the first result of electrolysis is the disappearance of protrusions from the surface of the active anode, that is, the anode surface is polished. This method is used in the *electropolishing* of surfaces.

This is by no means a complete list of the applications of electrolysis in modern technology.

## 22 Galvanic Cells and Storage Batteries

### 22-1 Transformation of Chemical Energy Into Electric Energy

When a metal plate is immersed in a solution containing ions, a potential difference is established between the metal and the solution, that is, the metal is electrified in the process. Let us discuss in more detail the processes that take

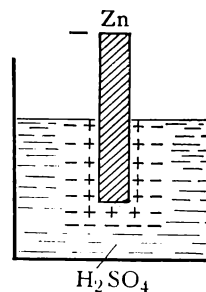
place when a zinc plate is immersed in a weak solution of sulphuric acid.

Since there are positive zinc ions on the surface of the plate, there will be negative  $\text{SO}_4^-$  ions in the solution near the plate, the hydrogen ions being forced away from the plate into the solution. The attraction of the zinc ions to the acid radical makes the  $\text{Zn}^{++}$  ions move from the surface of the plate into the solution. The result is that the plate acquires a negative charge and the solution a positive charge.

Since the zinc ions in the solution are attracted at the same time to both the  $\text{SO}_4^-$  ions and the negatively charged plate, they remain quite close to the surface of the metal (Fig. 22.4). In this fashion the zinc ions and the acid radical ions form a double layer close to the plate's surface, and a potential difference is established between the solution and the metal. If the plate is made of pure zinc, its dissolution will soon stop, prevented by the negative charge of the plate and the positive charge of the solution.

Hence, the contact of a solution with a metal results in an electric field possessing a definite energy. The formation of such a field involves the transformation of chemical energy into electric energy. It has been established by experiment that the potential difference between the solution and the metal depends on the nature of the metal and of the solution. This means that, if two plates of different metals are immersed into a solution, there should be a potential difference between them capable of creating an electric current. In this case the energy of the current is produced at the expense of chemical energy. It is possible to produce large amounts of energy in this way.

Fig. 22.1 Double ionic layer at zinc plate.

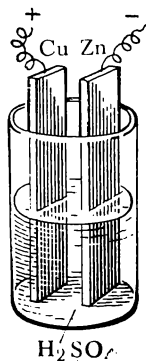


## 22-2 Galvanic Cells

An electric power source in which electric energy is produced at the expense of chemical energy is called a *galvanic* (or *voltaic*) cell. The invention of the first galvanic cell by the Italian physicist Alessandro Volta (1745-1827) at the end of the eighteenth century made it possible to produce continuous current and to investigate its regularities. We now know that the essential condition for the conversion of chemical energy into electric energy is the availability of two different conductors immersed in a solution with ionic conductivity.

The *simple voltaic cell* (Fig. 22.2), invented by Volta, consists of a weak aqueous solution of sulphuric acid into

Fig. 22.2 Simple voltaic cell.



which a copper plate and a zinc plate are immersed, the positive charge accumulating on the copper and the negative on the zinc. The potential difference between the plates (the emf of the cell) is approximately 1.1 V. The emf of a galvanic cell is independent of both the dimensions of the plates and the volume of the solution and is determined exclusively by the chemical processes taking place inside the operating cell. Let us consider these processes.

When an external circuit is connected to the cell's terminals, electrons from the zinc plate, the negative pole, start flowing from it to the copper plate, the positive pole. The departure of electrons from the zinc plate disturbs the equilibrium between it and the solution. The zinc ions concentrated in the double layer move away from the plate and new ions from the plate take their place. The decrease in the positive charge of the copper plate, in turn, enables new hydrogen ions to reach it. Upon contact with the plate these ions receive electrons and turn into gaseous hydrogen. Accordingly, in the process of operation of the simple voltaic cell hydrogen is liberated at its positive electrode and zinc dissolves at its negative electrode.

Experience has shown that the simple voltaic cell has a crucial deficiency: its emf drops rapidly and it stops producing electricity. Let us study this phenomenon more closely.

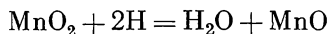
### 22-3 Polarization of Galvanic Cells and Its Reduction

In an operating simple voltaic cell gaseous hydrogen covers the copper plate and obstructs the discharge of the hydrogen ions. This leads to the concentration of positive ions at the copper electrode, these ions repelling other hydrogen ions and thus diminishing the current in the cell. The process is called *polarization*. In a simple voltaic cell the polarization of the copper electrode is due to the gaseous hydrogen liberated at it. Polarization can be said to produce a back emf in the cell, which reduces the current. Another name for this back emf is *polarization emf*. The principal cause of polarization in cells of other types is also the liberation of gases, mainly of hydrogen. Note that electrolysis, too, is accompanied by polarization, except in the cases where the anode dissolves in the electrolyte.

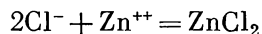
To prevent polarization a substance that reacts with the liberated gases should be introduced into the electrolyte.

Such a substance is called a *depolarizer*, the term for a cell containing a depolarizer being *nonpolarizable cell*. Such cells are adequately stable in operation and are widely used in practice. One such cell is the *Leclanché cell* (Fig. 22.3).

In the Leclanché cell a zinc plate serves as the negative electrode and a graphite rod as the positive. An aqueous ammonium chloride solution serves as the electrolyte, with manganese dioxide ( $\text{MnO}_2$ ) mixed with graphite powder and pressed around the graphite rod ( $\text{MnO}_2$  is used as the depolarizer). The hydrogen liberated at the positive electrode reacts with  $\text{MnO}_2$  according to the equation



and thus is not liberated in the form of gas at the graphite rod. Chlorine should be liberated at the negative electrode, but the chlorine ions react with zinc according to the equation



that is, zinc chloride is formed at the negative electrode. The emf of a Leclanché cell is 1.5 V. In dry cells, which are modified Leclanché cells the electrolyte is in the form of a paste made of flour and ammonium chloride.

## 22-4 Storage Batteries

It has been established that in some cases the polarization of the electrodes of a galvanic cell can remain in effect for a long time after the current in the electrolyte has ceased. This phenomenon is utilized in secondary cells — *accumulators*. Let us look into their principle of operation.

Lead electrodes immersed in a cell containing sulphuric acid are coated with lead sulphate. Since the chemical processes at both electrodes are identical, the potential difference between them is zero.

Connect the cell into a circuit (Fig. 22.4). The current flowing through it produces lead dioxide on the anode and pure lead on the cathode. Since the plates are now different, there is a potential difference between them. When the circuit is disconnected, this potential difference is equal to the polarization emf. Should we now connect an incandescent lamp instead of the battery it would burn. This means that the device retained its polarization emf and became itself a source of electric energy (Fig. 22.5). After

Fig. 22.3 Leclanché cell.

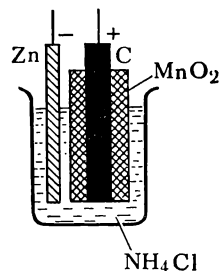


Fig. 22.4 Circuit for charging accumulators.

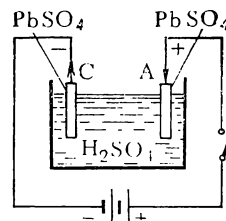
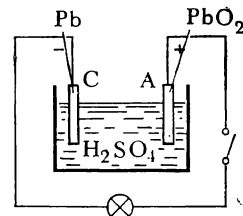


Fig. 22.5 Circuit for discharging accumulators.



some period of time both plates will again be coated with lead sulphate and the current will cease.

Now we can connect the accumulator to the power source again and repeat all the processes described above; this means that the processes in this device are reversible. Hence, our device accumulates electric energy when a current provided by an external source flows through it, and gives up this energy when it itself provides a current in another circuit. A battery of such secondary cells is termed a *storage battery*. Passing a current from an external power source through the accumulator is termed *charging*, and its operation as a power source is termed *discharging*.

The characteristic parameters of an accumulator are its efficiency, capacity and emf. The *efficiency* of an accumulator is the ratio of the energy given up in discharge to the energy spent on charging

$$\eta = W_{\text{dch}}/W_{\text{ch}} \quad (22.4)$$

The *capacity* of an accumulator is the maximum amount of electricity that can pass through a circuit in the process of the accumulator's discharge. The common unit of accumulator capacity is *ampere-hour*:  $1 \text{ Ah} = 3600 \text{ C}$ .

The most common accumulators in use are the acid and the alkali types. The lead-acid accumulator has already been described. Its emf is about 2 V and its efficiency is about 80 per cent. The emf of an alkali accumulator is about 1.3 V and its efficiency is not more than 60 per cent. Yet, it has some advantages over the lead-acid type: occasional short-circuiting does not destroy it, it is lighter, and it does not exude harmful vapours and gases.

### 22-5 Galvanic Cells and Storage Batteries in Modern Life

The electricity produced in galvanic cells is comparatively weak. Therefore such cells are used mainly in low-power applications, such as telephony and telegraphy. Galvanic and storage batteries are widely used in portable radio and television sets, in pocket calculators and in various mechanical toys. The use of storage batteries is widespread in transportation. They are used in power trucks carrying loads at railway stations and factories. In automobiles storage batteries are used for starting up the engine and for parking lights. Storage batteries power submarines.

The widespread use of automobiles in modern society is the cause of considerable harmful air pollution. This

is the reason for the appearance of experimental prototypes of electric cars, which may in the future prove a substitute for cars powered by internal combustion engines. Modern electric cars are powered by storage batteries.

Note that great harm is caused by *local cells* spontaneously appearing in metal structures at points of contact of different metals. The part of the electrolyte in such local cells is usually played by water, which nearly always contains a large enough number of ions.

In moist air a water film covers metal objects and local cells start producing currents which corrode the metals. Such currents are termed *spurious galvanic currents* and their effect on metals is termed *corrosion*.

## Electric Current in Gases and in Vacuum

## 23

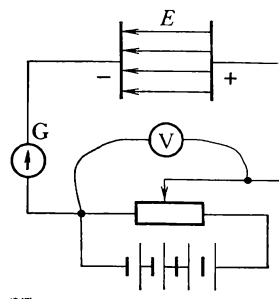
### 23-1 Ionization of a Gas

All gases in normal conditions are good insulators, but in a small enclosed space any gas, including air, can be made to conduct electricity. To this end mobile charge carriers must be created in them, that is, the gas molecules must be ionized.

The following experiment serves to clear up the point. Take a large parallel plate capacitor (Fig. 23.1), with the plates two to three centimetres apart, and connect it to a power source of several thousand volts. A sensitive galvanometer  $G$  connected into the circuit indicates the absence of current despite the presence of an electric field between the plates. This means that either there are no free carriers in the air between the plates or that their number is so small that the galvanometer is unable to respond to their motion. We shall see below that the latter is true.

Place a burning candle in the space between the plates or direct a beam of X rays at it. The galvanometer pointer will move, that is, there is a current in the circuit. This means that some molecules of air have been *ionized* (mobile charge carriers have been generated). If the ionizer is withdrawn, the current dies down rapidly because the air between the plates again turns into a dielectric. Such experiments proved that high temperature, X rays, ultraviolet rays,  $\alpha$ -rays, etc. may serve as ionizing agents for gas molecules.

Fig. 23.1 Circuit for studying current in gases at atmospheric pressure (arrows indicate field direction).





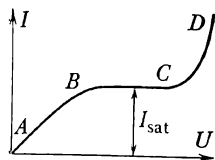
In a gas, in addition to ionization, there is always a reverse process taking place—the *recombination* of ions, that is, the formation of neutral molecules from ions. (Why in the case of the continuous operation of an ionizing agent and in the absence of an electric field a dynamic equilibrium is established in the gas?)

The number of charge carriers per unit volume of gas is greater the higher the intensity of the ionizer, that is, the greater the number of ions it generates each second. Each singly ionized gas molecule loses one of its valence electrons. Some of these attach themselves to neutral gas molecules, creating negative ions, and some remain free. Therefore the charge carriers in an ionized gas are free electrons and ions (positive and negative). Hence the conductivity of an ionized gas is partly ionic and partly electronic. (Consider the difference between the ionization of an electrolyte and a gas.)

### 23-2 Dependence of Current on Voltage

Raising the voltage  $U$  across the capacitor's plates and measuring the current  $I$  (Fig. 23.1) with a galvanometer, one can obtain the dependence of the current flowing in a gas on the voltage (the volt-ampere characteristic of the gas discharge gap) for a constant ionizing agent (Fig. 23.2). It can be seen from the plot that the current in a gas obeys Ohm's law only in the low-voltage range. Let us look for an explanation for this.

Fig. 23.2 Graph of current in gas versus voltage at atmospheric pressure.



In the case of a small voltage between the plates the charge carriers acted upon by the electric field move at low speeds and in most cases manage to recombine before they reach the plates. When the voltage is increased, the velocity of the ions moving in the electric field increases and the probability of their recombination diminishes. Because of this the number of ions reaching the plates per unit of time and being neutralized on them grows, that is, the current increases (segment  $AB$  in Fig. 23.2). Hence, in this case the increase in current is due to the decrease in the recombination of mobile charge carriers in the gas.

If the voltage across the plates is further increased, the time will come when the recombination of charge carriers will essentially end and the current will reach its maximum value  $I_{\text{sat}}$ , which will no longer depend on the voltage (segment  $BC$  in Fig. 23.2). Indeed, in the absence of recombination all the carriers generated by the ionizing agent will reach the plates and for this reason a further increase in voltage fails to bring about an increase in current. Note

that in order to increase maximum current one has to increase the intensity of the ionizer. The term for a current independent of voltage is *saturation current*.

Hence Ohm's law is not applicable to the part *BC* of the characteristic. Note in addition that the ions discharging at the plates again associate into molecules of the same gas from which they were produced. This means that a current flowing in gas does not, of itself, entail chemical processes and that therefore Faraday's laws are not applicable to it.

At high voltages, when the field strength in the gap rises to several tens of thousands volts per centimetre, free electrons are accelerated by the electric field over their mean free path to such high energies that on striking the gas molecules they tear electrons away from the molecules. The term for this ionization is *impact ionization*. Impact ionization increases the number of charge carriers in the electrode gap and a further increase in voltage is accompanied by a sharp increase in current (section *CD* in Fig. 23.2).

### 23-3 Electric Discharge Through Gases at Atmospheric Pressure

A discharge through a gas which takes place only in the presence of an external ionizing agent is termed *semi-self-maintained*. Another term for it is *quiet discharge*, since it can be detected only with the aid of instruments. A discharge in a gas which takes place without an external ionizer is termed *self-maintained*.

It was demonstrated above that the charge carriers in a gas are free electrons and ions. But when a current flows in a gas the ions are discharged at the electrodes and turn into neutral atoms and molecules and electrons are absorbed by the positive electrode. Besides, a fraction of the carriers vanishes as a result of recombination. Therefore, to maintain the current in the gas there should be a mechanism to compensate for the constant depletion of charge carriers. In the case of semi-self-maintained discharge, as we already know, this job is done by the external ionizing agent. In case of a self-maintained discharge it is done by the current itself.

There are several mechanisms for the generation of new charge carriers in a gas. One of them is the impact ionization mentioned in the preceding section. Let us discuss the conditions required for impact ionization in more detail.

We recall that to move an electron (charge  $e$ ) from a point of the field with potential  $\phi$  to some point outside

the field work must be performed against the forces of the field equal to  $W = \varphi e$ . Therefore to ionize a gas molecule work  $W_{\text{ion}}$  must be performed. This can be expressed by the relation

$$W_{\text{ion}} = \varphi_{\text{ion}} e \quad (23.1)$$

The potential  $\varphi_{\text{ion}}$  is the *ionization potential* of an atom or molecule. Its magnitude depends on the nature of the atom or molecule.

To be able to ionize a gas molecule in collision with it an electron should possess, before collision, a kinetic energy K.E. greater than or equal to the ionization work  $W_{\text{ion}}$ :

$$\text{K. E.} \geq W_{\text{ion}} \quad (23.2)$$

The electron must gain this energy in travelling a distance in the external electric field in the gas equal to its mean free path,  $\lambda$ , since after each collision the electron loses its speed and then starts gaining it again. The force acting on the electron is  $Ee$  (where  $E$  is the field strength) and the mean free path of the electron is  $\lambda$ , therefore

$$\text{K.E.} = Ee\lambda, \text{ or } mv^2/2 = Ee\lambda \quad (23.3)$$

where  $m$  is the electron mass, and  $v$  is its speed before collision with a molecule.

Since at atmospheric pressure the mean free path of electrons is small, high field strengths are required for impact ionization of a gas to take place. Hence, at atmospheric pressure impact ionization takes place only at high voltages.

Gradually increasing the voltage across the electrodes (Fig. 23.4), we attain a field strength high enough for the appearance of impact ionization. The number of collisions which result in impact ionization are at first small, but grow with voltage.

Secondary electrons born in the process of impact ionization are also accelerated by the field and take part in ionization. Eventually, at a certain voltage most electrons, before disappearing from the current, will each be the cause of the ionization of at least one (on the average) gas molecule and of the creation of a new free electron. In such conditions not only will the discharge in a gas be able to maintain itself, but the process of impact ionization may develop into an avalanche, in which case the avalanche carrier multiplication will result in a rapid increase in the current and in an *electric breakdown* of the gas. The few free electrons which are always present in a gas are enough to initiate such a self-maintained discharge.

Note that ions colliding with gas molecules also cause impact ionization. However, they are much less effective than electrons, mainly because their mean free path (and consequently their kinetic energy) is substantially less than that of the electron.

We will now consider other mechanisms of charge carrier generation operating in a self-maintained discharge.

When the negative electrode is operating at high temperatures, it becomes a source of *thermionic emission*. This can substantially increase the number of free electrons in the gas. Positive ions are then attracted to the negative electrode and, if their kinetic energy is high enough, can knock electrons out of it at the moment of impact. The term for this phenomenon is *secondary emission*.

High voltages are required to produce secondary emission in a cold cathode at atmospheric pressure. However,

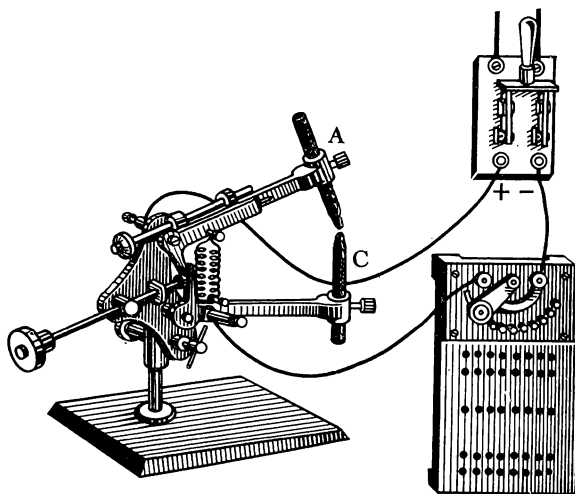


Fig. 23.3 Carbon arc—direct current.

in the case of a hot cathode self-maintained discharge is possible at low voltages. An example of such a discharge is the carbon arc, discovered in 1802 by the Russian physicist V. V. Petrov (1761-1834).

Connect two carbon rods (A and C in Fig. 23.3) in series with a rheostat to a power source with a voltage of about 100 V. By bringing the rods into contact we close the circuit and see that they become red-hot since the resistance in this part of the circuit is the greatest. Next separate the rods. The current continues to pass through the air and a bright arc appears between the rods. The ends glow even hotter and radiate a blinding light. A crater is formed on

the anode, and the cathode point sharpens (Fig. 23.4). Note that metal electrodes can also be used to produce an arc. High temperatures of the rods (about  $4000^{\circ}\text{C}$  for the anode and about  $3000^{\circ}\text{C}$  for the cathode) are sustained by the continuous bombardment of their surfaces by charged air particles. The current in the arc may be as high as several tens or even hundreds of amperes (it is limited by the rheostat).

An important part in the generation of mobile charge carriers is played by the radiation of the arc itself which ionizes the gas. The resistance of the gas gap in an arc discharge depends on the current, as is the case with all types of self-maintained discharge. Because of this Ohm's law is non-applicable to it. Thus, *arc discharge* is the term for an electric discharge in a gas involving a red-hot cathode. The electric arc is widely used in industry, for instance, in electric arc furnaces, in the electrolytic method of aluminium production, for electric welding, as a powerful light source in searchlights, etc.

*Spark discharge* is the term for the intermittent discharge in a gas which takes place at voltage high enough for the formation of an avalanche. A high current at the moment of a spark discharge causes the voltage across the electrodes to drop and this, in turn, causes the discharge to die down. After some time the necessary voltage builds up across the electrodes again and another spark appears. These discharges follow in a rapid sequence which the eye perceives as one continuous spark in the form of zigzag luminous lines connecting the electrodes. Note that if the power of the source is high enough, a spark discharge may turn into an arc discharge.

The spark is a thin branching line of highly ionized gas. Owing to its high conductivity it carries very strong currents. The gas in the spark is heated to high temperatures and shines brightly. A rapid increase in temperature causes sound effects.

A dramatic example of a spark discharge occurring in nature is lightning. The voltage between the Earth and a cloud in a thunderstorm can be as high as several hundred million volts, while the current in lightning exceeds 100 000 A. The explanation of the forked pattern of lightning is that the discharge passes through regions of air with the smallest resistivity. In a gas such regions form a random pattern.

*Brush* and *crown discharges* develop in a gas when impact ionization takes place only in the vicinity of the electrodes or wires where the field strength is at its greatest, not in

the entire space of the field. The avalanches die out as they reach regions of lower field strength. Such discharges take place at voltages somewhat lower than the voltages required for a spark discharge.

Brush discharge can be obtained if one electrode is made in the shape of a disk and the other a point. It has the appearance of a luminous beam connecting the point with the disk.

Crown discharge appears around high tension wires. It is accompanied by a glow and characteristic crackling. In the course of this discharge the ions near the wire come in contact with it and are discharged, causing losses of the power transmitted through the wire. Therefore a crown around high-tension lines is to be avoided. Crown discharge is employed to advantage in electric filters for cleaning furnace gases which are apt to pollute air with coal particles.

### 23-4 Electric Discharge Through Gases at Low Pressure

Comparing formulae (23.2) and (23.3), we see that for a self-maintained discharge in a gas to be possible the following relationship should be fulfilled:

$$Ee\lambda \geq W_{\text{ion}} \quad (23.4)$$

This means that by increasing the mean free path of the electrons we can obtain a self-maintained gas discharge at lower field strengths, that is, at a lower voltage. Hence, the conductivity of a gas should increase with its rarefaction. The following experiment is proof of this conclusion.

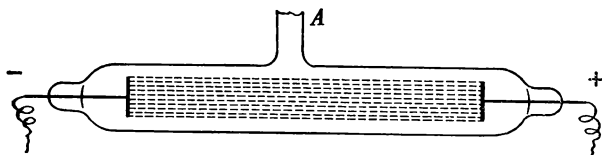


Fig. 23.5 Discharge tube with gas at low pressure.

Take a glass tube with two sealed-in electrodes with an open tube *A* for evacuating the air out of the tube (Fig. 23.5). Next connect the electrodes to a power source with a voltage of several thousand volts and start evacuating the air out of the tube. As the pressure drops the air in the tube starts to glow. This means that a self-maintained discharge has started in the tube. The free electrons in the tube are the secondary electrons generated at the cathode by ions striking its surface.

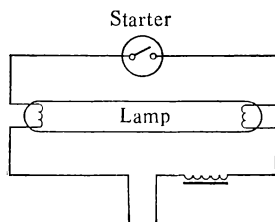
The nature of air luminescence changes with rarefaction. First purple lines appear between the electrodes and then the entire air in the tube begins to glow with a rosy light. Filling the tube with different gases and passing current through them one observes each rarefied gas to have its own characteristic glow. For instance, argon glows with a blue light and neon with a red light.

The term for a discharge in a rarefied gas accompanied by luminescence is *glow discharge*. The luminescence of a glow discharge is not very bright. Heat liberated in the gas in the course of a glow discharge is not great and the glowing gas stays cold. Glow discharge is evident in the fluorescent lamps widely used for advertising signs.

A modern method of internal lighting is *daylight lamps*, fluorescent lamps with walls coated with a special, *lumino-phor*, coating. The luminescence of the luminophor is excited by the radiation of the rarefied gas itself, the result of current flowing through the gas-filled lamp. The luminophor absorbs mainly invisible radiation and emits visible radiation of a spectral composition similar to that of the visible radiation from the Sun. Daylight lamps contain rarefied inert gas and mercury vapour. A schematic diagram of a circuit containing a daylight lamp is shown in Fig. 23.6.

With the circuit closed, a current flows through the *starter* and heats the electrodes inside the lamp, initiating thermionic emission from them. This starts a self-maintained discharge in the lamp, the high temperature of the electrodes being maintained by the current flowing in the lamp. The starter is then automatically disconnected.

Fig. 23.6 Circuit diagram for "daylight" lamp.



### 23-5 Radiation and Absorption of Energy by an Atom

The difference in the luminescence of various gases in glow discharge is explained by Bohr's theory, developed in 1913. Even before this theory it had been established that molecules and atoms are responsible for visible radiation.

Niels Bohr suggested that an electron in an atom cannot move in arbitrary orbits, since there are only *allowed* orbits, each corresponding to a specific energy of the atom. The rule for the selection of these orbits will be discussed in Section 38-15. With an electron on one of the allowed orbits the atom's energy remains constant. The minimum energy of the atom corresponds to the electron moving in the orbit nearest to the nucleus. This state of the atom is termed *normal*, or *ground*. It can persist for an indefinite time.

With the electron in any other orbit the state of the atom is said to be *excited*, the atom's energy being the higher the more distant is the orbit occupied by the electron from the nucleus. An atom cannot remain long in an excited state and after a short time (about  $10^{-8}$  s) the electron jumps over to an orbit nearer to the nucleus. Figure 23.7 is a schematic representation of the possible electron transitions from one orbit to another in a hydrogen atom.

To effect the transition of an atom from the normal to an excited state a quite definite portion of energy termed a *quantum*, must be supplied to it. The greater the quantum absorbed by the electron the more distant from the nucleus is its new orbit.

Hence, the transition of an atom from the ground to the excited state takes place only as a result of the action of external forces capable of supplying it with the necessary amount of energy. Such action may be a collision of the atom with an electron, an ion, or another atom. The absorption of radiation energy also brings about an atom's transition to an excited state. When subsequently the atom returns to the ground state, it itself emits electromagnetic radiation of a definite wavelength, which carries away energy liberated by the atom. The greater the energy liberated by the atom on each occasion the shorter the electromagnetic wave (see Section 31-3).

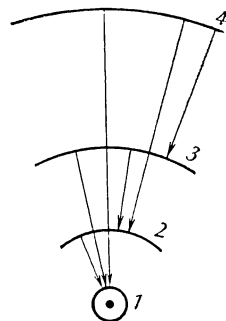
Since visible light is electromagnetic radiation in a definite wavelength range (the colour is determined by the wavelength), the transition of atoms from different excited states to the ground state results in radiation of different colours. Since the variations in the energy of atoms of a single type in transitions from a definite excited state to the ground state are identical, the set of colours radiated by all such atoms is the same. It is a fact that the set of colours radiated by atoms of a specific chemical element is different from that radiated by atoms of another element. All these questions will be treated at length in Sections 38-15 and 38-16.

The collisions of atoms with electrons and ions in the course of a glow discharge excite the atoms in the gas and it begins to radiate light of a colour depending on the nature of the gas.

### 23-6 Cathode Rays

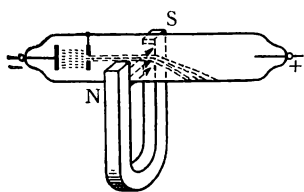
It was stated in Section 23-4 that rarefaction of a gas reduces its electric resistance. However, at very low pressures (of the order of a thousandth of a mmHg) the opposite

Fig. 23.7 Circular orbits according to Bohr's theory of hydrogen atom showing energy changes.





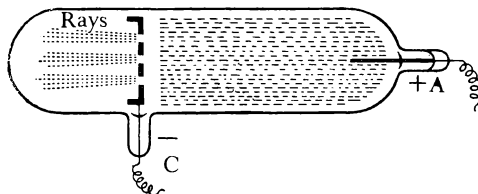
**Fig. 23.8** Deflection of cathode rays in magnetic field.



is true. At such pressures the luminescence of the gas disappears almost completely and the glass of the tube opposite the cathode shines with a greenish light. The explanation of this phenomenon is as follows.

With a small number of gas molecules remaining in the tube the collisions between the electrons and the molecules become rare and this explains the disappearance of gas luminescence. In a high vacuum most electrons and ions cover the whole interelectrode distance without colliding. The speed of the positive ions in the tube before they strike the cathode is determined by the voltage across the electrodes, a typical value being of the order of  $10^5$  m/s. In striking the cathode the ions knock electrons out of it, and the electrons move at right angles to the cathode. They are accelerated in the interelectrode field and attain maximum speed at the anode (typical value is  $10^7$  m/s). The electrons which miss the anode strike the glass and excite its luminescence. An electron beam flying in a high-vacuum tube at right angles to the cathode is termed a *cathode ray*. Note that collisions between the electrons and the gas molecules remaining in the tube are essential, since only in this case will ions responsible for the secondary electron emission from the cathode be produced. If there are not enough gas molecules in the tube, the current will cease, since an absolute vacuum does not conduct current (it is an ideal dielectric). If cathode rays consist of an electron flux, they should be deflected in electric and magnetic fields in the manner peculiar to moving negative charges (see Section 25-9). Such deflection can indeed be observed (Fig. 23.8).

**Fig. 23.9** Anode rays in space behind cathode.



Cathode rays excite the luminescence of many substances, they are also the cause of mechanical and thermal effects. Note that it was the study of cathode rays that led to the discovery of the electron and to the determination of its charge and mass.

An interesting phenomenon can be observed in a tube containing highly rarefied gas. If holes are drilled in the tube's cathode, it becomes possible to observe the motion of positive gas ions in the space behind the cathode

(Fig. 23.9). When voltage is applied to such a tube, rays termed *positive*, or *canal*, appear behind the cathode. They constitute a flux of positive ions of the gas remaining in the tube. The properties of canal rays are in many ways similar to those of cathode rays, but their deflection in magnetic and electric fields is much smaller and opposite in direction to that of electrons.

### 23-7 Plasma

We recall that a gas without mobile charge carriers (free electrons or ions) is a dielectric and that an ionized gas is a conductor, despite the fact that on the whole it is electrically neutral since it contains equal numbers of positive and negative charge carriers. The term for a gas a substantial proportion of whose atoms or molecules is ionized is *plasma*.

Hence, plasma is matter in a state in which it is on the whole electrically neutral but contains free positive and negative charge carriers in equal numbers. If plasma also contains neutral atoms or molecules, it is called *partially ionized*. When all its molecules or atoms are ionized the term is *completely ionized plasma*.

At temperatures above 20 000 K any substance is a completely ionized plasma. This is the most widespread state of matter in nature. The Sun and other stars, in which almost the entire substance of the universe is concentrated, are gigantic coagulations of high-temperature plasma.

Partially ionized plasma makes up the upper layers of the atmosphere (the ionosphere). Such plasma, but in an extremely rarefied state, is dispersed in outer space. A gas carrying electric current is also an example of partially ionized plasma.

### 23-8 Electric Current in Vacuum

It was mentioned above that an absolute vacuum is an ideal dielectric. To make current flow in a high vacuum, charge carriers, electrons, should be introduced into it. This can be achieved with the aid of thermionic emission by placing in a vacuum a wire that can be connected to an external circuit.

An example of such a device is depicted in Fig. 20.2. When a vacuum tube is used, the electrons from the red-hot filament fly out into the vacuum. An electric field established between the filament C and the electrode A makes the

electrons move to electrode A, thereby closing the circuit, and so a current flows in the vacuum. In this case the electrons move unhindered in the vacuum and gain kinetic energy at the expense of the work performed by the forces of the field. If the voltage across the electrodes in Fig. 20.2 is  $U$ , the work of the field in moving an electron from electrode A to electrode C will be

$$W = Ue$$

Since this work turns into the kinetic energy of the electron we obtain

$$\text{K.E.} = Ue \quad \text{or} \quad mv^2/2 = Ue \quad (23.5)$$

Here  $m$  is the mass,  $v$  the speed, and  $e$  the electron charge. The term for  $U$  in this case is the *accelerating voltage*. The mass of an electron is very small, which makes it easy to control the motion of the electrons in a vacuum.

### 23-9 The Diode

The control of the motion of electrons in a vacuum by means of an electric field is the principle of operation of vacuum tubes whose external appearance is similar to that shown in Fig. 20.2.

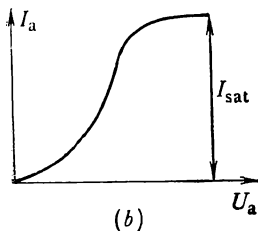
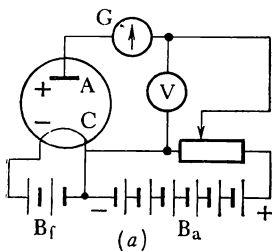
The term for the simplest vacuum tube with two electrodes is *diode*. One of its electrodes is a tungsten wire whose ends pass out of the tube. This makes it possible to heat the wire by passing current from a battery  $B_f$  (Fig. 23.10a). The space inside the tube is evacuated to  $10^{-6}$ – $10^{-7}$  mmHg.

With the wire C serving as the *cathode* which is made red-hot, thermionic emission takes place and electrons appear in the tube. The second electrode, A, serves as the *anode* (or *plate*). It can be connected to the cathode by means of an anode battery  $B_a$ . Note that the anode has one wire leading out of the tube, and so a hot-cathode diode has a total of three leads (electrodes) to be connected to an external circuit.

When the anode is disconnected and the cathode is red-hot, electrons create a negative *space charge*. If the temperature of the cathode remains constant, there will be a dynamic equilibrium at its surface between the electrons evaporating from the cathode and condensing on it. (Explain why.) This means that the number of electrons in the space charge of the tube remains constant. To increase the space charge the cathode temperature should be raised.

Now connect the anode battery so that the anode is at

Fig. 23.10 (a) Basic diode circuit; (b) anode characteristic of diode at constant filament current.



the negative terminal and the cathode at the positive. In this case the electric field inside the tube will displace the electrons towards the cathode, so that there will be no current in the anode circuit. This can be established with the aid of a galvanometer,  $G$  whose pointer stays at rest in this case.

Let us see what will happen if the positive terminal of the battery  $B_a$  is connected to the anode and its negative terminal to the cathode (see Fig. 23.10a). In this case the field will drive the electrons to the anode, that is, there will be a current flowing in the tube and the galvanometer's pointer will move.

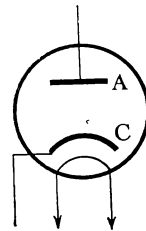
Hence, electronic tubes are remarkable for their property of conducting current in only one direction. This is the basis for the most important function of the diode—rectification of alternating current. Vacuum diodes used for this purpose are termed *rectifiers*.

A volt-ampere characteristic measured at a constant cathode temperature is termed the *anode characteristic curve* (Fig. 23.10b). At first the current increases with the voltage, this being due to the dispersal of the space charge, with the result that a smaller number of electrons from it condense on the cathode. As the voltage is further increased nothing is left of the space charge and all electrons leaving the cathode reach the anode. Thus no further increase in the voltage increases the current, that is, the current attains its saturation value  $I_{\text{sat}}$ . Its magnitude rises with cathode temperature. It follows from this that Ohm's law is not applicable to vacuum tubes.

To increase emissivity thermionic cathodes are coated with a film of barium and strontium oxides. This produces a sharp decline in the work function and raises the emission by several orders of magnitude. The term for such a cathode is *heater cathode*. Note that the maximum current for which the diode is designed is usually much less than its saturation current.

Tubes used in mains-operated electronic equipment have cathodes heated indirectly by alternating current (Fig. 23.11).

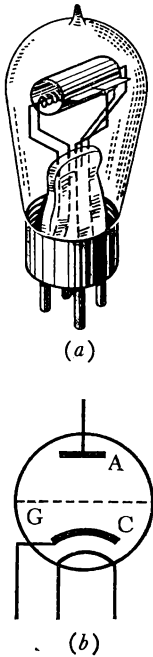
Fig. 23.11 Circuit symbol for diode with indirectly heated cathode.



### 23-10 The Triode

A very important property of the vacuum tube is that it is practically without inertia. This is because electrons are the lightest charge carriers and the current in a tube is able to follow very rapid changes in electrode voltage.

**Fig. 23.12** (a) Triode with directly heated cathode; (b) circuit symbol of triode with anode A, grid G and indirectly heated cathode C.

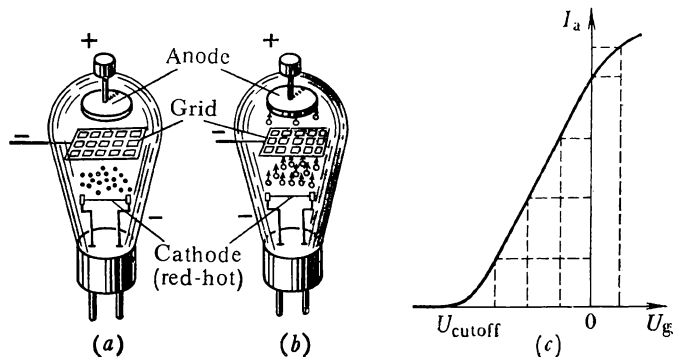


An effective means of controlling current in an electronic tube is the placing of an additional electrode between the cathode and the anode. This electrode is called a *grid*. The grid is placed close to the cathode and because of that even a small voltage applied between the grid and the cathode establishes a strong electric field in the space between them, thus greatly affecting the anode current.

Usually the grid is a wire spiral wound around the cathode quite close to it. The anode in this case is a solid cylindrical surface surrounding the grid and the cathode. The term for a three-electrode tube with a grid is *triode* (Fig. 23.12). With zero grid bias (zero grid-cathode voltage) there is a current that is set up by the anode voltage in the tube. At a negative bias (when the grid potential is below that of the cathode) a field retarding the progress of the free electrons towards the anode appears between the grid and the cathode. Only electrons possessing sufficient kinetic energy are able to overcome the retarding field and fly through the grid to the anode, the rest will be forced back to the cathode. The result is a sharp decline in the anode current, the current stopping altogether when the bias is sufficiently negative (Fig. 23.13a).

The grid-cathode voltage at which the current ceases to flow is termed *cut-off bias* ( $U_{\text{cutoff}}$ ). If the grid voltage is above the cut-off bias, there will be a current flowing in the anode circuit (Fig. 23.13b), small variations in grid voltage causing large variations in anode current. Figure 23.13c shows the dependence of the anode current on the grid voltage (*grid characteristic* of the tube). This makes it possible to use the electron tube for amplification of electric signals. Figure 23.14 shows a schematic diagram of an amplifier. An alternating voltage applied to the tube's

**Fig. 23.13** (a) At large negative grid bias ( $|U_g| > |U_{\text{cutoff}}|$ ) electrons remain close to cathode and there is no current ( $I_a = 0$ ); (b) at  $|U_g| < |U_{\text{cutoff}}|$  electrons reach anode and anode current flows in triode's circuit; (c) dependence of anode current  $I_a$  on grid voltage  $U_g$ .



grid  $U_{\text{sig}}$  results in substantial variations in the anode current  $I_a$  (Fig. 23.15). The voltage across the load resistance varies in proportion to the anode current:  $U_L = I_a R_L$ , those variations being tens or even hundreds of times greater than the variations in grid voltage.

The anode current in the tube varies in proportion to the grid voltage, provided the grid voltage  $U_g$  remains within the linear part of the grid characteristic (see Fig. 23.13c). With a positive potential on the grid the increments in the anode current are no longer proportional to the increments in  $U_g$  (i.e. the electric signals will be distorted in the process of amplification). Besides, some electrons will reach the grid, creating a current in the grid circuit and thus loading the signal source. This also causes distortion of the signal. Hence, the grid should always be

Fig. 23.14 Circuit diagram of an amplifier of electric signals.

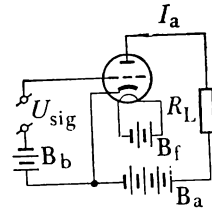
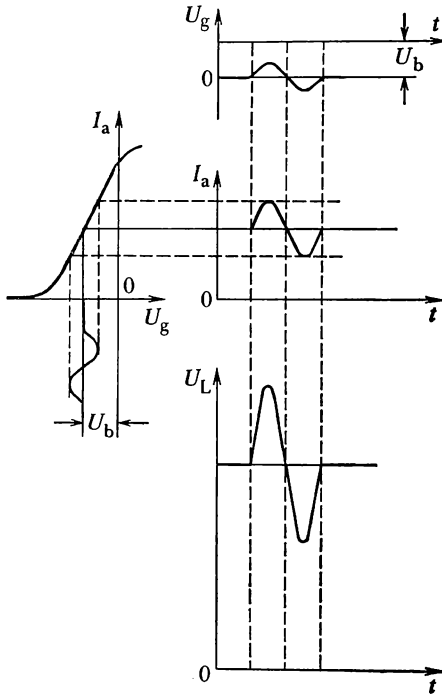


Fig. 23.15 Grid voltage  $U_g$ , anode current  $I_a$  and voltage across load resistance  $U_L = I_a R_L$  versus time.



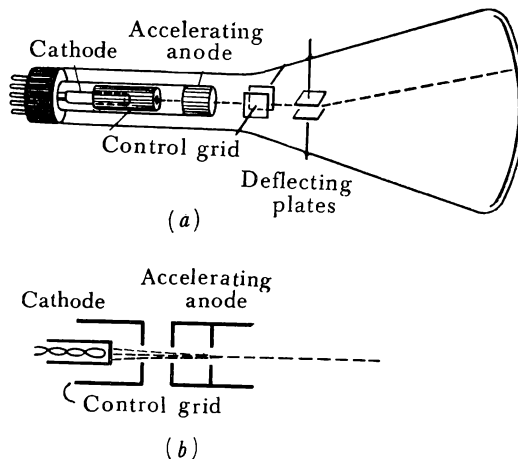
at a negative potential with respect to the cathode. This is done by means of a bias source connected into the grid circuit ( $B_b$  in Fig. 23.14) which establishes a negative bias  $U_b$  on the grid (see Fig. 23.15).

To improve the characteristics of amplifier tubes additional grids are introduced in them. A tube with two grids is termed a *tetrode* and a tube with three grids a *pentode*.

### 23-11 The Cathode-Ray Tube

The *cathode-ray tube* is used to obtain images by projecting an electron beam onto a screen. It is a vital part of oscillographs, television sets, radar displays and various other electronic equipment. The tube is a hermetically sealed evacuated glass flask with a wide face (Fig. 23.16a). The neck of the flask contains the *electron gun* (Fig. 23.16b)

**Fig. 23.16** (a) Construction of cathode-ray tube using electrostatic deflection; (b) the arrangement of electrodes in electron gun.



which produces the electron beam. The electron gun consists of a heated cathode and a *control grid* in the shape of a cylinder, acting in a manner similar to the grid in the triode.

Thermionic emission takes place at the heated cathode. The electrons fly towards the anode, passing on the way through an opening in the control grid. The control grid controls the number of electrons flowing to the anode and helps to focus them into a narrow beam, the term for which is *electron beam*. The anode is made of several disks with holes in them. These disks are placed inside a hollow metal cylinder. Such an arrangement also helps to focus the beam on the head of the flask serving as screen.

A voltage of several thousand volts is applied between the anode and the cathode of the tube. The field between the cathode and anode accelerates the electrons to high

speeds. Therefore when they have flown the length of the flask they strike the screen coated with a luminophor. The latter lights up and a bright spot appears on it.

A possible method of controlling the displacement of the electron beam in the tube is by applying an additional lateral electric field with the aid of *deflecting plates*. To this end two mutually perpendicular pairs of such plates are arranged in the tube (see Fig. 23.16a). The field in one of these capacitors deflects the electron beam in the horizontal direction and the field in the other in the vertical direction. This method makes it possible to displace the bright spot to any point of the screen. Note that the position of the electron beam can also be controlled with the aid of magnetic fields set up by two coils. This type of deflection is used in tubes for television sets.

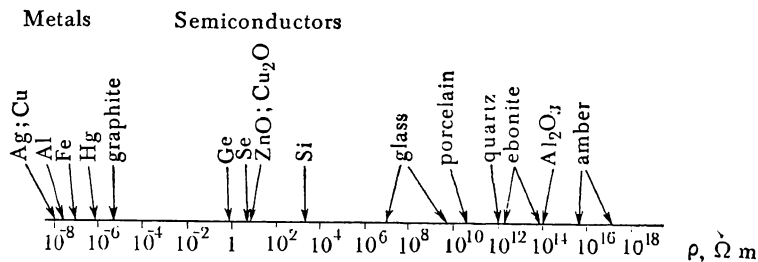
## Electric Current in Semiconductors

# 24

### 24-1 Conductors, Dielectrics and Semiconductors

Until recently all substances from the point of view of their electric properties were divided into conductors and dielectrics. Such a division was quite sensible since there is a great difference between the electrical conductivities of these substances (Fig. 24.1).

The specific resistance of conductors lies within the range  $10^{-5}$  to  $10^{-8} \Omega \cdot \text{m}$ , and of dielectrics  $10^{10}$  to  $10^{16} \Omega \cdot \text{m}$ . These



**Fig. 24.1** Specific resistance of some materials at room temperature.

figures illustrate the extent of the gap between the values of the specific resistances of conductors and dielectrics.

Subsequent research into the electric conductivity of substances led to the discovery of materials whose electric



conductivities lie between the values for conductors and dielectrics (see Fig. 24.1). These substances were named *semiconductors*. They include first of all the elements of group IV of the Mendeleev Periodic Table: germanium, silicon, compounds of group III elements and group V elements ( $A_{III}B_V$  compounds),  $A_{III}B_{VI}$  compounds, group IV compounds such as silicon carbide; many oxides; some group VI elements (selenium, tellurium); and many other substances. The specific resistance of semiconductors lies in the range of  $10^8$  to  $10^{-5} \Omega \cdot m$ .

Note that the specific resistance of various substances is affected by the presence of imperfections (impurity atoms or lattice defects). The presence of impurities has little effect on the concentration of free charge carriers in metals, but appreciably affects their mobility. Impurities in metals, if they are not integrated in a regular lattice, increase specific resistance, because the presence of even a few irregularities in otherwise perfect metal lattice substantially increases the possibility of an electron being scattered, that is, to lose its energy.

The electrons in a dielectric are strongly bound to the atoms of the regular lattice. Where there are imperfections, this bond is loosened and the electron in it may be freed. Hence, the conductivity of a dielectric is mainly due to imperfections contained in it, the specific resistance of a dielectric decreasing with the increase in impurity concentration.

In semiconductors, as in dielectrics, the presence of impurities generally reduces resistivity. By an appropriate choice of the type and concentration of impurities it is possible to change the specific resistance of a semiconductor in the desired direction. This is why impurity semiconductors are widely used in modern technology.

It may be interesting to compare the temperature dependence of the specific resistance of various substances. Recall that the specific resistance of metals rises with temperature, decreasing when the metal is cooled and becoming zero in the superconducting state. The specific resistance of dielectrics, however, decreases with temperature. Because great energy is required to tear an electron away from an atom of a dielectric, solid dielectrics usually melt or break up before attaining appreciable conductivity.

The energy required to tear an electron from the atom of a semiconductor is much less than in the case of a dielectric. Therefore, while the number of free charge carriers in an intrinsic semiconductor increases with the rise in temperature in the same way as in a dielectric, it can be-

come conductive before melting or breaking up. The specific resistance of intrinsic semiconductors rises with the decrease in temperature and in the low temperature range can be as high as that of dielectrics. Semiconductors do not become superconductive.

It has been established by experiment that it is not only temperature that affects the specific resistance of semiconductors. Irradiating a semiconductor with light substantially reduces its resistance, because radiation brings with it energy sufficient to generate free charge carriers (see Section 38-10).

Thus, the conductivity of semiconductors depends on the temperature and on illuminance. These properties of semiconductors are very important for practical applications.

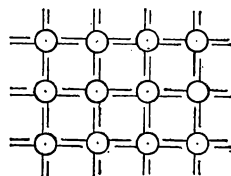
## 24-2 Pure (Intrinsic) Semiconductors

Using germanium and silicon as examples, let us consider the process of charge carrier generation in intrinsic semiconductors in more detail. The atoms of these elements have four valence electrons in their outer shell. In the solid state these substances crystallize in the diamond lattice in which every atom has four nearest neighbours (see Fig. 13.11). In this lattice neighbouring atoms form covalent bonds, that is, two neighbouring atoms share their valence electrons (one each), the electrons forming an electron pair (Section 13-3). A schematic diagram of the covalent bond is shown in Fig. 24.2 with the space lattice symbolically represented by a two-dimensional one.

At low temperatures all the electrons in a semiconductor are tied to their atoms. Such a crystal has no free charge carriers and therefore is a dielectric. When the temperature of such a crystal is gradually raised, some of the electrons may gain enough excess energy (at the expense of the energy of random motion) to break their bonds with the atoms. It is the appearance of such electrons that makes the semiconductor crystal conductive. The energy required to tear an electron away from the lattice ions is less for germanium than for silicon. Hence, at an equal temperature the specific resistance of germanium is much less than that of silicon (at 20°C  $\rho_{\text{Ge}} = 0.6 \, \Omega \cdot \text{m}$  and  $\rho_{\text{Si}} = 2 \times 10^3 \, \Omega \cdot \text{m}$ ).

The transition of an electron to the free state leaves a vacant state in the bond structure of the lattice, the accepted term for which is *hole*. Since before the departure of the electron the respective locality of the lattice was neutral, its departure gives to the hole a positive charge. The neigh-

Fig. 24.2 Covalent bonds between atoms of Group IV of Mendeleev Periodic Table.



bouring atoms in the semiconductor continuously exchange electrons, so a hole in one bond can be filled by an electron from another bond, a hole then being created in the latter.

In this way positively charged holes in a semiconductor take part in random motion in much the same way as the electrons. Holes in a semiconductor are thus regarded as mobile charge carriers. Indeed, if in the absence of an electric field in the semiconductor the holes move at random, the application of an external field causes them to move in the direction of the field, that is, sets up a current.

Hence, heating a semiconductor results in the *generation* of electron-hole mobile carrier pairs. Free electrons and holes moving at random in the semiconductor may meet. In that case, other conditions being favourable, the free electron may fill the vacant bond, with the result that two free charge carriers disappear at the same time, through the *recombination* of an electron-hole pair. The average time a generated free carrier exists in a semiconductor before it recombines, that is, its life-time, depends on many factors and may be as high as several milliseconds and as low as  $10^{-8}$  s.

At constant temperature and in the absence of a current there is a state of dynamic equilibrium between the processes of generation and recombination of electron-hole pairs and corresponding specific concentrations of mobile charge carriers of both types. (Explain why an increase in temperature leads to an increase in the number of mobile charge carriers simultaneously existing in a semiconductor and to a decrease in the specific resistance of an intrinsic semiconductor.)

Note that the numbers of free electrons and holes in an intrinsic semiconductor are always equal. Therefore the conductivity of intrinsic semiconductors is partially electron-type and partially hole-type. The term for such bipolar conductivity is *intrinsic conductivity*.

If an intrinsic semiconductor is connected into a circuit, a current will flow in it, the electrons moving from the negative terminal to the positive and the holes in the opposite direction. One should not forget that in the latter case the current is in fact carried by bonded electrons moving from bond to bond under the influence of the field.

The temperature coefficient of the resistance of intrinsic semiconductors is many orders of magnitude greater than that of metals (and opposite in sign). Advantage is taken of this fact in the design of devices which automatically switch off circuits to protect them from excessive temperatures. A semiconductor device whose resistance at normal

temperatures is high is connected into a circuit containing a bell or current control device. When the temperature exceeds the limit, the resistance of the semiconductor device drops and a current appears in the signal circuit, actuating the bell or cutting off the current responsible for the overheating. Such semiconductor devices are termed *thermistors*. They are used to measure temperatures and radiation (by its thermal effect). Very small in size, they can be used to measure temperature in a small space.

### 24-3 Impurity (Extrinsic) Semiconductors

By doping pure semiconductors with special impurities semiconductors with electron- or hole-type conductivities can be synthesized.

Add as an impurity about  $10^{-5}$  per cent some element of group V of the Mendeleev periodic table, for example arsenic to molten pure germanium. The melt will crystallize into a normal germanium lattice, but some sites will contain arsenic atoms instead of germanium atoms (Fig. 24.3). Four valence electrons of an arsenic atom form covalent bonds with neighbouring germanium atoms, with the fifth electron so loosely bonded to the arsenic atom that quite a small energy is required to tear it away.

For this reason all arsenic atoms in germanium are ionized at normal temperatures. The positive arsenic ions are localized in the lattice and are unable to move in an external field, the free electrons (one per impurity atom) acting as mobile charge carriers.

Such a crystal has electron-type conductivity, also termed *n-type conductivity* (*n* for negative). The crystal itself is termed *n-type semiconductor*. The impurity which donates free electrons to the semiconductor is termed *donor*, or *n-type*, impurity.

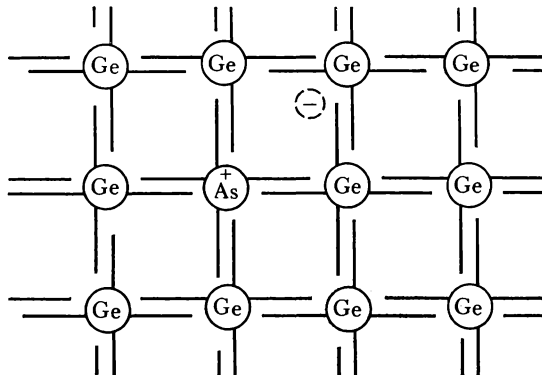
If germanium is doped with an element of group III of the Mendeleev Periodic Table, for instance, indium, the atoms of which each have three valence electrons, each such atom will be able to establish only three bonds with the neighbouring germanium atoms. In order to form a bond with its fourth neighbour in the germanium lattice the indium atom must borrow an electron from one of its neighbours and turn into a negative ion, creating a hole in place of the borrowed electron which thereafter wanders at random in the crystal (Fig. 24.4).

A germanium crystal doped with a group III element has hole-type conductivity, also termed *p-type conductivity* (*p*

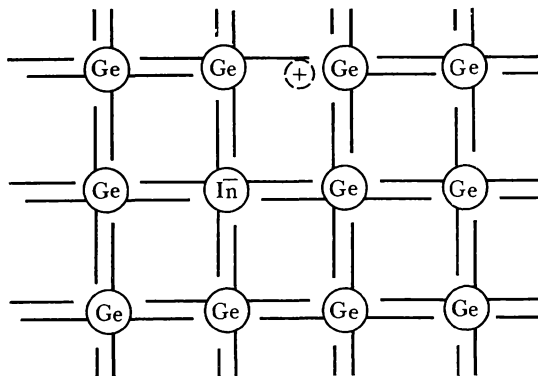
for positive). The impurity responsible for such conductivity is termed *acceptor*, or *p-type*, impurity.

Note that in semiconductors some carrier pair generation already takes place in normal conditions, so that an extrinsic semiconductor also contains besides the majority carriers, a certain proportion of minority carriers. At low

**Fig. 24.3** *n*-type impurity atom (As) having lost one of its electrons became a positive ion-localized positive charge and mobile (free) electron has been formed.



**Fig. 24.4** *p*-type impurity atom (In) having trapped electron became negative ion-localized negative charge and mobile hole (lacking electron) has been formed.



temperatures this proportion is quite small. However, at high temperatures, when substantial generation of electron-hole pairs takes place, the conductivity of the semiconductor approaches the bipolar type. Hence, extrinsic semiconductors retain their mainly unipolar conductivity only at temperatures below the temperature of transition to intrinsic conductivity.

### 24-4 P-N Junction

Imagine a germanium crystal, one half of which contains donor impurity and the other acceptor impurity. The boundary separating the  $n$ -type region from the  $p$ -type region is termed  $p$ - $n$  junction. Consider the properties of this junction.

Suppose that both regions have just been brought into contact (although in fact they are part of the same single crystal). If this were to happen, the electrons would start moving from the  $n$ -region, where they are in abundance, to the  $p$ -type, where there is a shortage of them. The same would be true for the holes, but in the opposite direction. Such diffusion of electrons and holes would continue until their concentrations in both parts became equal, were it not accompanied by charge transport. However as a result of the latter the  $n$ -region acquires a positive charge and the  $p$ -region a negative one, that is, a contact potential difference is established between the  $p$  and the  $n$ -regions.

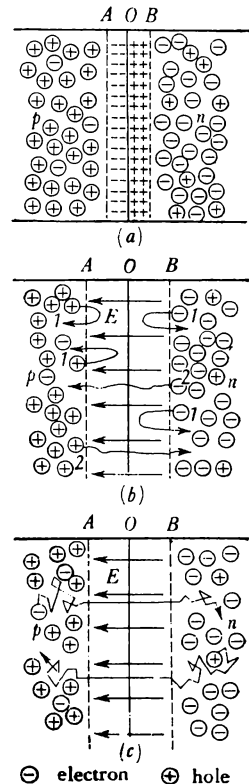
In the transition region  $AB$  (Fig. 24.5a) of the  $p$ - $n$  junction an electric field is established obstructing any further diffusion of charge carriers across it. Only holes and electrons possessing a sufficiently high kinetic energy are able to overcome the opposition of the field and cross the transition region  $AB$  (Fig. 24.5b). On the other hand, this field encourages the transition of the minority carriers in opposite directions to those of the majority carriers, of holes from the  $n$ -region and of electrons from the  $p$ -region. Indeed, should a free electron moving at random in the  $p$ -region cross the boundary  $A$  of the transition region (Fig. 24.5c), it would be forced by the field to the  $n$ -region, the same being the fate of the holes from the  $n$ -region.

In equilibrium a potential difference is established across the layer  $AB$  (about one volt) which provides for the compensation of the diffusional flux of the holes by the hole current from the  $n$ -region set up by the field in the transition layer  $AB$ . The electron currents flowing from both regions are compensated in the same way. Accordingly, the net hole and electron currents both become zero.

There are practically no mobile charge carriers in the transition layer  $AB$  since they penetrate it at high speed. Only the localized ions of the acceptor impurity remain in the  $AO$  region of the transition layer, with the ions of the donor impurity in the  $BO$  region. The charges of the  $p$ - and  $n$ -regions are concentrated in these regions, other parts of the crystal being electrically neutral.

The transition layer, when depleted of mobile charge carriers, has despite its small width (of the order of

Fig. 24.5  $p$ - $n$  junction: (a) layer  $AB$  depleted of mobile charge carriers is formed between  $p$ - and  $n$ -regions and entire field is concentrated in this layer; region  $AO$  contains localized ions of  $p$ -type impurity and region  $OB$  localized ions of  $n$ -type impurity; (b) formation of diffusional flux of majority carriers flowing across junction (1, electrons and holes unable to overcome opposition of field; 2, electrons and holes of sufficient energy to overcome opposition of field); (c) current of minority carriers across junction under action of field in region  $AB$ .



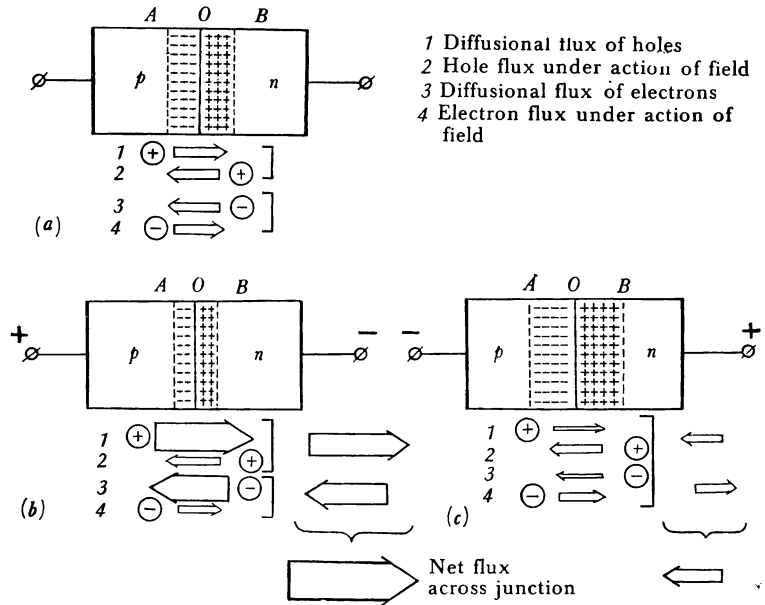
$1\text{ }\mu\text{m}$  ( $10^{-6}\text{ m}$ )) a very great resistance as compared with other parts of the crystal. Therefore, when a crystal containing a  $p$ - $n$  junction is connected into a circuit, practically the entire voltage applied to the crystal is concentrated across the  $p$ - $n$  junction.

### 24-5 The Semiconductor Diode

Let us turn our attention to the current flowing through a crystal containing a  $p$ - $n$  junction. Such a crystal is depicted schematically in Fig. 24.6.

In the absence of external voltage (Fig. 24.6a) all mobile carrier fluxes across the junction are compensated and the current is zero.

**Fig. 24.6** (a) Current is zero because net fluxes of holes and electrons through junction  $AB$  are zero; (b) field in junction is reduced—diffusional fluxes of majority carriers prevail over reverse fluxes of minority carriers and substantial forward current flows across junction; (c) field in junction is increased—minority carrier fluxes prevail over diffusional fluxes of majority carriers and small reverse flux flows across junction.



Connect the crystal into a circuit so that the external field is directed against the field of the  $p$ - $n$  junction (Fig. 24.6b). The field in the  $p$ - $n$  junction will be weakened and the majority carriers (holes from the  $p$ -region and electrons from the  $n$ -region) will rush across the junction, the fluxes of minority carriers remaining practically unaffected. The result is a large current flowing across the  $p$ - $n$  junction. The term for this direction of the voltage and current

is *forward*. The current rises very rapidly with the voltage (Fig. 24.7), Ohm's law here being quite out of place.

Now apply a voltage of opposite polarity to the crystal (Fig. 24.6c). In this case the external field coincides in direction with the contact field. The external field augments the field of the  $p$ - $n$  junction, decreasing the diffusional majority carrier fluxes across it. The minority carrier currents are of about the same order of magnitude as in the absence of the external field, but this time, uncompensated by the majority carrier fluxes, they combine to make

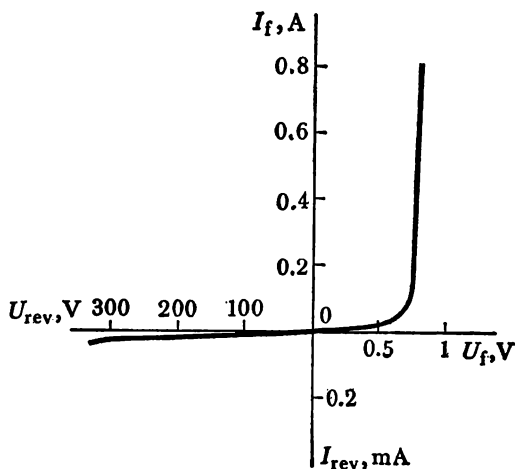


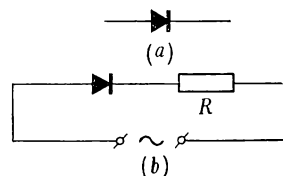
Fig. 24.7 Volt-ampere characteristic of silicon medium-power diode (scales for forward and reverse currents are different).

up a small current across the junction (remember that the concentrations of minority carriers in the  $p$ - and  $n$ -regions are small). The term for such direction of voltage current is *reverse*.

The over-all result is that the forward current across the  $p$ - $n$  junction is millions of times greater than the reverse current for a voltage of equal magnitude (Fig. 24.7). This means that the  $p$ - $n$  junction operates like a *rectifier*, that is, it conducts current in one direction (the junction is open) but not in the opposite direction (the junction is closed). Accordingly, if the crystal containing a  $p$ - $n$  junction is connected into a circuit in series with a load resistance  $R$  (Fig. 24.8), the current in that circuit will flow only in one direction, that is, it will be rectified (see Section 30-8). (Figure 24.7 depicts the volt-ampere characteristic of a medium-power silicon rectifier diode.)

*Semiconductor rectifier diodes* feature high efficiency (up to 98 per cent), small size and long life. Their deficiencies

Fig. 24.8 (a) Circuit symbol for diode; (b) diode in series with load.





include the deterioration of their operational parameters at elevated temperatures. It was mentioned above that the reverse current in the  $p$ - $n$  junction is due to the minority carriers, whose concentrations at normal temperatures are small, but rise rapidly at higher temperatures because of increased electron-hole pair generation. On account of this the reverse current of semiconductor diodes rises rapidly with the increase in temperature: silicon diodes stop rectifying current at a temperature of about 200 °C, the maximum temperature for germanium diodes being even less.

Note in addition that a semiconductor diode should not be connected into a circuit without a load resistance (for instance, connected directly to a capacitor with a high capacitance). If the load in Fig. 24.8*b* is short-circuited the entire voltage will be applied to the diode. Because of the small forward resistance of a semiconductor diode at high currents (several ohms, or even less for power diodes; see Fig. 24.7), the application of a forward voltage will in this case produce a current so large that it will put the diode out of action.

## 24-6 The Transistor

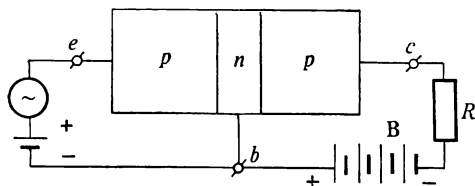
The properties of the  $p$ - $n$  junction are utilized in semiconductor amplifiers of electric signals.

Three-electrode semiconductor devices used for amplifying or generating voltages or currents are termed *transistors*. A schematic diagram of a transistor is depicted in Fig. 24.9. In the  $p$ - $n$ - $p$  transistor a narrow (of the order of 1  $\mu$ m)  $n$ -region separates two  $p$ -regions of a single crystal, each region having its own contact electrodes  $e$ ,  $b$  and  $c$  for connection into a circuit. The transistor in the diagram is seen to have two  $p$ - $n$  junctions. The external voltage to the left-hand  $p$ - $n$  junction is applied by means of the electrodes  $e$  and  $b$  and to the right-hand junction by means of the electrodes  $b$  and  $c$ .

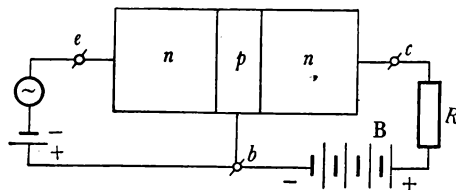
The left-hand  $p$ -region of the transistor has a  $p$ -type impurity concentration several thousand times higher than that of the  $n$ -type impurity in the  $n$ -region, the ratio of the hole concentration in the  $p$ -region to the electron concentration in the  $n$ -region being somewhat less. Therefore, when the left-hand junction is connected in the forward direction, the forward current across the junction will mainly (about 99 per cent) consist of the diffusional current of holes from the  $p$ -region.

Let us see how such a transistor operates as a voltage amplifier. Connect a load resistance (see Fig. 24.9) to the right-hand junction and apply to it a large (tens of volts) reverse voltage. Since the junction is closed, a very small reverse current should be flowing across it, not enough to cause a noticeable voltage drop across the resistor  $R$ .

Apply a small forward voltage to the left-hand junction. A direct current consisting almost entirely of holes diffusing from the  $p$ -region to the  $n$ -region will start flowing across it. The  $n$ -region is so narrow that most holes entering it from the  $p$ -region cross it in a time much shorter than their life-time in that region. Therefore, they are unable to recombine in it before they reach the right-hand junction. Holes in the  $n$ -region are minority carriers and on entering the right-hand junction are driven by its field across it to



**Fig. 24.9** Connection of transistor as voltage amplifier. Small ac input voltage appears with amplified amplitude across the resistor  $R$ .



**Fig. 24.10** Circuit for  $n$ - $p$ - $n$  transistor.

the right-hand  $p$ -region. Hence, with the left-hand junction open a current almost equal to that across the left-hand junction will flow across the right-hand junction, instead of a small reverse current, and a substantial voltage  $U = IR$ , equal to the emf of battery  $B$ , will be received by the resistor  $R$ .

Since the forward resistance of the  $p$ - $n$  junction is small (see Fig. 24.7), the current in the load will vary greatly for small variations in the voltage across the left-hand junction: a change in it of some decimal fractions of a volt causes a change in the voltage across  $R$  of several tens of volts.

The operation of an  $n$ - $p$ - $n$  transistor shown in Fig. 24.10 is in principle identical to that of a  $p$ - $n$ - $p$  transistor, with

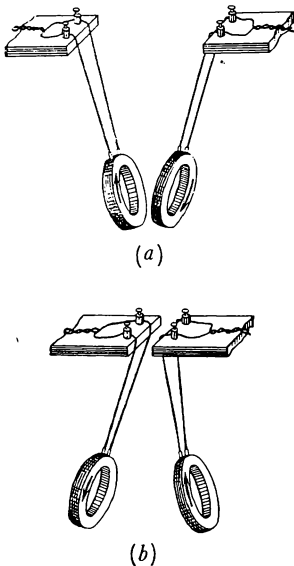
the exception that the electrode polarities are reversed and the current in the middle region is carried mainly by electrons.

Transistors have a long life, are very economical and are small in size. They are widely used in radioelectronics: in amplifiers, in radio and television sets, in computers and in other apparatus. The properties of the transistor are of special value for electronic equipment, aircraft and missiles.

Note that *emitter* is the term for the region which supplies the carriers to the middle region, termed *base*, the term *collector* describes the region collecting the carriers—their designations are *e*, *b*, *c*, respectively (see Figs. 24.9 and 24.10).

## 25 Electromagnetism

Fig. 25.1 (a) Conductors carrying currents in same direction attract each other; (b) conductors carrying currents in opposite directions repel each other.



### 25-1 Interaction of Currents

We have already discussed the interaction of charges. Let us now find out whether there is any interaction between conductors carrying currents.

Take two identical coils of wire and suspend them so that they can be connected into a circuit and so that they have a common axis (Fig. 25.1). If we pass currents through the coils in the same direction, we see that they are attracted to each other (Fig. 25.1a). If one of the currents is reversed, the coils are repelled (Fig. 25.1b). The same sort of interaction can also be observed between two straight conductors arranged in parallel.

Hence, currents going in the same direction are attracted, while those going in opposite directions are repelled. Therefore, when two conductors are carrying currents and are at some distance from each other, there is an interaction between them which cannot be explained by an electric field, since such conductors remain neutral. This means that around every current-carrying conductor there is a field of a different nature than that of an electric field, since it has been shown in experiment that this new field does not act on static charges.

## 25-2 Magnetic Field as a Special Form of Matter

We will agree to term the field by which the interaction of currents takes place the *magnetic field*. It has been established by experiment that a magnetic field is always the result of moving charges and acts only on such charges. Hence, to detect a magnetic field in some region of space a current-carrying conductor, or some other moving charges, should be introduced into it. The first to discover a magnetic field around current-carrying conductors in experiment was the Danish physicist Hans Christian Oersted (1777-1851) in 1820.

The magnetic fields of various currents may, when superimposed, either augment or weaken each other. Let us demonstrate this in an experiment. If two identical coils are put together and currents are passed through them in opposite directions (Fig. 25.2*a*, left), their combined field will become so weak that it will have no noticeable effect on the third current-carrying coil. This explains the absence of a magnetic field around a bifilar line of two wires carrying currents in opposite directions. But if the directions of the currents in both coils coincide, they will act on the third coil with a much stronger force (Fig. 25.2*b*) than in the case of the experiment described in the preceding section. Hence, the superposition of unidirectional currents amplifies the magnetic field, the superposition of opposite currents weakens it.

If before the experiment the coils are arranged so that their axes do not coincide, then when the current is switched on the coils will, of themselves, turn so that the currents flowing in them coincide in direction, with the coils being attracted. This increases the magnetic field in the space around the coils.

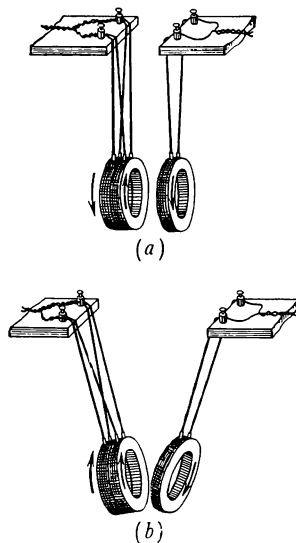
## 25-3 Magnets

The properties of magnets justify the assumption that they are surrounded by a magnetic field of a definite direction. One is entitled to ask the following questions.

(1) Can it be assumed that the magnetic fields around a magnet and around a current-carrying conductor are identical in nature, that is, that the field in both cases is a magnetic field?

(2) If both fields are magnetic, how can this be reconciled with the statement that only electric current can establish a magnetic field?

Fig. 25.2 (a) Magnetic fields of currents flowing in opposite directions weaken each other; (b) magnetic fields of currents flowing in same direction amplify each other.



Oersted's experiments have proved that the nature of the magnetic field of a current-carrying conductor is the same as that of a magnet.

Ampere answered the second question. According to Ampere's theory there are molecular currents (microcurrents) inside a magnet similar to a current flowing in a closed circuit. It has been subsequently established that such microcurrents are due to electrons moving inside the atoms. This means that, in general, there should be a magnetic field around every molecule (atom). Consequently every substance must possess some magnetic properties determined by the nature of the motion of its electrons and molecules and by the mutual arrangement of its molecules, that is, by the peculiar character of the internal structure of the substance.

Iron has proved to be the most interesting substance in this respect. The fields of its atoms are fairly strong and, if arranged so as to augment each other, are able to establish a magnetic field around an iron body. Such bodies are *magnets*. When the molecules in a body are arranged at random, their fields are mutually compensated and there is no magnetic field around the body. By placing such an iron body in a magnetic field, for instance inside a current-carrying coil, the body can be magnetized because its atoms are arranged in a regular order by the external field.

Bodies made of special steels retain their magnetization after being withdrawn from the magnetic field—they become *permanent magnets*. A permanent magnet attracts bodies containing iron. The greater force of attraction is at the ends of the magnet termed the *magnetic poles*. A small magnet of an elongated shape, able to rotate on a point, is called a *magnetic needle*. In the absence of interference the magnetic needle automatically turns so that one of its ends points to the north and the other to the south. The end of the needle pointing northward is called the *north-seeking pole* or *N pole*, and the opposite end is called the *south-seeking pole* or the *S pole*.

Since a definite direction can be attributed to every point of a magnetic field (recall the compass) it has been agreed to define the direction of a field at any one point as the direction indicated by the N pole of a magnetic needle placed at this point. It has been established by experiment that like poles of magnets are repelled and unlike are attracted.

### 25-4 Magnetic Lines of Force

It is a well-known fact that a magnetic needle turns about its axis when placed in a magnetic field. This means that magnetic forces forming a force couple act on its ends (Fig. 25.3a). When the needle stops, the forces act in the direction coinciding with that of the needle (Fig. 25.3b;

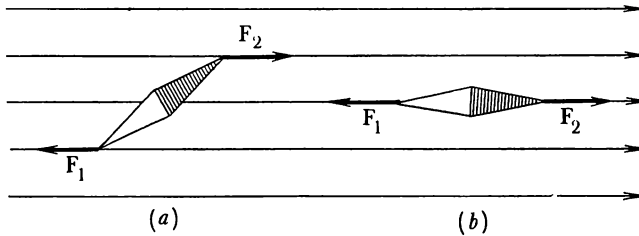


Fig. 25.3 Force couple turns magnetic needle in magnetic field.

explain why). This means that magnetic needles can be used to find the direction of the lines along which magnetic forces act.

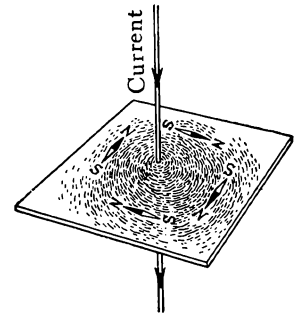
A magnetic field is schematically represented in diagrams by *magnetic lines of force*, also called *lines of magnetic induction*. This is the term for a line so drawn that a tangent to it at any point indicates the direction of a small magnetic needle placed in the field.

In practice a picture of lines of magnetic induction can easily be obtained with the aid of steel filings, since every steel particle is magnetized in the magnetic field and turns into a tiny magnetic needle, which arranges itself in the direction of an induction line. Figure 25.4 depicts the magnetic field of a rectilinear current in a plane perpendicular to the conductor, obtained with the aid of filings and several magnetic needles. The conventional direction of lines of induction coincides with the direction indicated by the N poles of the needles, that is, in this case it is clockwise if viewed from above (in the direction of the current).

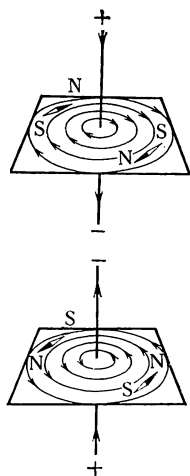
Only one line of induction passes through any point in space. Thus they never intersect.

The lines of magnetic induction in Fig. 25.4 are seen to be closed, that is, they have neither a beginning nor an end, and always circle around the current-carrying conductor. This is a very important property of lines of magnetic induction. We recall that electric field strength lines start and end at electric charges (or at infinity). A field whose lines of induction are always closed is termed *solenoidal*. In contrast to the potential field of electric charges the magnetic field is always solenoidal.

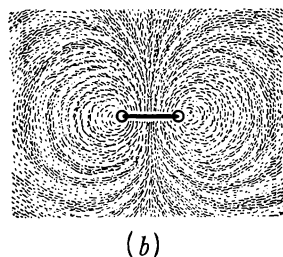
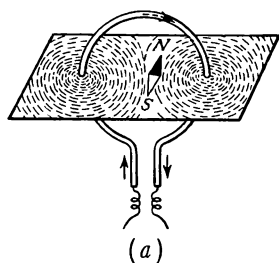
Fig. 25.4 Magnetic field of straight current made visible due to steel filings and magnetic needles.



**Fig. 25.5** Magnetic lines of force of straight current.



**Fig. 25.6** (a) Magnetic field of circular current; (b) magnetic field of circular current made visible by steel filings (viewed from above).



The obvious conclusion from the above is that a magnetic field and an electric current always exist together. In nature there is never a magnetic field without an electric current and never an electric current without a magnetic field.

### 25-5 Magnetic Fields in Some Simple Cases

The magnetic field of a current-carrying conductor is determined by the magnitude and direction of the current, as well as by the shape of the conductor.

Magnetic fields of straight conductors carrying opposite currents are depicted in Fig. 25.5. It can be seen that the only difference between the fields is the direction of their induction lines. The magnetic field of a straight current is in the shape of concentric circles lying in any plane perpendicular to the conductor. The direction of the lines of magnetic induction is determined by the *right-hand screw rule*: if the translational motion of the screw coincides with the direction of the current in the conductor, the direction of the screw's rotation will indicate the direction of the lines of magnetic induction.

The magnetic field of a *current loop* is depicted in Fig. 25.6. The position of the magnetic needle in Fig. 25.6a indicates the direction of the lines of magnetic induction. It is determined by the right-hand screw rule.

Note that the right-hand screw rule can also be given in another form: if the screw is rotated in the direction of the current flowing in the loop, the translational motion of the screw will indicate the direction of the induction lines inside the loop.

The magnetic field of a *solenoid*, a current-carrying coil, is depicted in Fig. 25.7. The induction lines of a solenoid are seen to be parallel to its axis inside the solenoid and to envelop it. As in the case of a current loop the direction of the induction lines for a solenoid can be found by applying the right-hand screw law.

### 25-6 Comparing Magnetic Properties of a Solenoid and a Permanent Magnet

Comparing the magnetic field of the permanent bar magnet depicted in Fig. 25.8 with the magnetic field of the solenoid shown in Fig. 25.7, one is bound to observe a similarity. The only difference is in the pattern of the

magnetic fields inside the solenoid and the magnet (the pattern of the induction lines inside a magnet is not visible).

The magnetic properties of a current-carrying solenoid and a bar magnet are practically identical. For instance, if a solenoid is suspended so that it can rotate freely in a

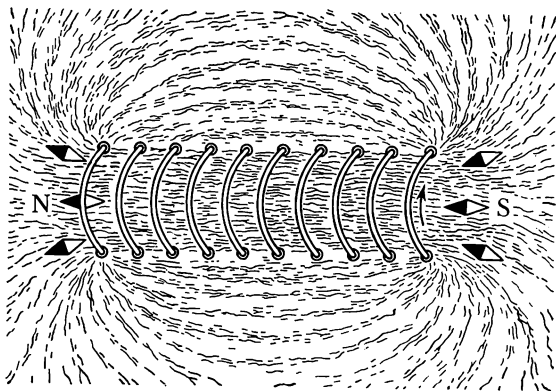


Fig. 25.7 Magnetic field of current-carrying coil (direction of current indicated by arrow) made visible due to steel filings and magnetic needles (field is uniform inside coil).

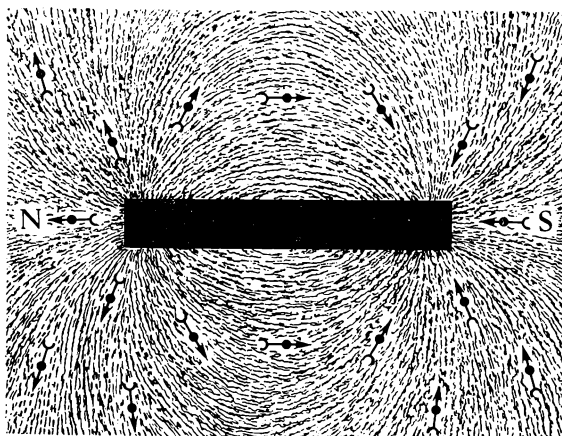
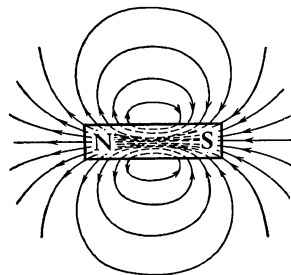


Fig. 25.8 Magnetic field of permanent magnet (arrows indicate direction of induction lines).

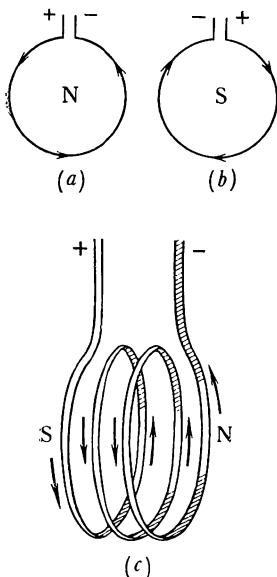
horizontal plane, it will automatically turn to point in the north-south direction. This agrees well with Ampere's idea of the field of a magnet being the result of molecular microcurrents. The above justifies the assertion that the induction lines of a magnet are closed, that is, that they continue inside it just like those of a solenoid (Fig. 25.9).

Since a magnet has poles, one concludes that a solenoid should also have them. Actually, if one end of a current-

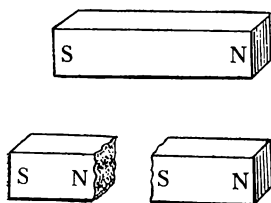




**Fig. 25.10** Magnetic poles of circular current, (a) and (b), and of current-carrying coil, (c).



**Fig. 25.11** If permanent magnet is broken in two, both parts become separate magnets.



carrying solenoid attracts the N pole of a magnet, its other end repels it. Having determined the direction of the induction lines with the aid of the right-hand screw rule, one can find the magnetic poles of a solenoid, as in the case of a magnet the induction lines leave the solenoid at its N pole and enter it at its S pole. The same method is used to determine the poles of a current loop (its poles are on the surface encompassed by the loop).

The above may serve to define the pole of a solenoid or a circular current: the surface around which the current flows counterclockwise is the N pole, while the surface around which it flows clockwise is the S pole (Fig. 25.10). The poles of a solenoid are shown in Fig. 25.7.

Note that when two observers look at the same loop from opposite sides, one of them sees the current flowing clockwise and the other counterclockwise. Therefore each current-carrying loop must have two different poles. Hence, magnetic poles exist only in pairs. It is impossible to have a single magnetic pole. Breaking a permanent magnet in two, we obtain two magnets, each with an N and an S pole (Fig. 25.11).

### 25-7 Interaction Between Parallel Currents. The Permeability of a Medium

In future we shall refer to the forces of interaction between currents and magnets as *magnetic forces*. Let us determine the factors influencing the force of interaction between two parallel currents (Fig. 25.12).

If the directions of the currents  $I_1$  and  $I_2$  flowing in the conductors coincide, the currents will attract each other with forces  $F_1$  and  $F_2$ . The existence of these forces is due to the fact that the second conductor is in the magnetic field of the first, this field being the source of force  $F_2$ . Obviously the force  $F_1$  is, in turn, the result of the magnetic field of the second conductor.

By varying the currents in the conductors and the distance  $a$  between them one can demonstrate that the force  $F$  acting on a section  $l$  of a long wire is directly proportional to the product of the currents and to length  $l$  and inversely proportional to the distance  $a$ :

$$F = K \frac{I_1 I_2 l}{a} \quad (25.1)$$

By changing the medium containing the conductors it can be established that force  $F$  is dependent on the medium.

Accordingly, the factor  $K$  depends both on the choice of the units of measurement and on the medium. Research into the problem has led to the conclusion that a medium can both augment and reduce the interaction between the currents as compared with a vacuum. We can agree to express the dependence of the force  $F$  on the medium with the aid of the quantity  $\mu_m$ . In this case the factor  $K$  can be presented in the form

$$K = k\mu_m \quad (25.2)$$

Here factor  $k$  in (25.2) depends only on the choice of units of measurement. The quantity  $\mu_m$  expressing the dependence of the force of interaction of electric currents on the medium is called the *permeability* of the medium. With account taken of (25.2) formula (25.1) can be written in the form

$$F = k \frac{\mu_m I_1 I_2 l}{a} \quad (25.3)$$

The value of  $k$  in the SI system is  $1/2\pi$ ; therefore, we have finally

$$F = \frac{\mu_m I_1 I_2 l}{2\pi a} \quad (25.4)$$

In practice it is more convenient to use relative permeability  $\mu$ . *Relative permeability* is the ratio of the force of interaction between currents in the specified medium to that in a vacuum. Denoting the force of interaction of currents in a vacuum by  $F_0$  and the permeability of a vacuum (called *permeability of free space*) by  $\mu_0$ , we obtain from (25.4)

$$F_0 = \frac{\mu_0 I_1 I_2 l}{2\pi a} \quad (25.4a)$$

Since it follows from the definition of  $\mu$  that  $\mu = F/F_0$ , dividing both sides of (25.4) by (25.4a), we obtain

$$\mu = \mu_m/\mu_0 \quad \text{or} \quad \mu_m = \mu\mu_0 \quad (25.5)$$

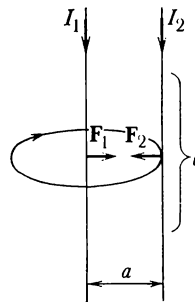
It should be noted that tables usually contain values for relative permeability  $\mu$ .

## 25-8 Definition of the Ampere

Formula (25.4a) is used in the International System of Units to define the unit of current, the ampere. With equal currents in the conductors it takes the form

$$F_0 = \frac{\mu_0 I^2 l}{2\pi a} \quad (25.4b)$$

**Fig. 25.12** Forces between parallel currents appear because each is in magnetic field of other.



It follows from (25.4b) that, if we place two parallel conductors carrying equal currents in a vacuum at a definite distance  $a$  and specify the force which should act on a unit length of each conductor, the magnitude of the current will be fully determined. This is used to define the ampere: an *ampere* is the current flowing in each of two infinite parallel conductors placed in a vacuum at a distance 1 m from each other which produces a force of  $2 \times 10^{-7}$  N on every meter of their length.

Note that in the SI system the term for the permeability of a vacuum is *magnetic constant*. Let us find its numerical value. It follows from (25.4b) that

$$\mu_0 = \frac{2\pi a F_0}{I^2 l}$$

Substituting the values corresponding to the definition of the ampere we obtain\*

$$\begin{aligned} \mu_0 &= \frac{2\pi \times 1 \text{ m} \times 2 \times 10^{-7} \text{ N}}{1 \text{ A}^2 \times 1 \text{ m}} = 4\pi \times 10^{-7} \frac{\text{N} \times \text{m}}{\text{A}^2 \times \text{m}} \\ &= 4\pi \times 10^{-7} \frac{\text{N}}{\text{A}^2} \end{aligned}$$

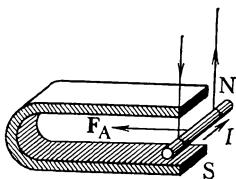
### 25-9 A Measure of the Strength of the Magnetic Field

Let us examine the factors affecting the magnitude of a magnetic force acting on a straight current-carrying conductor.

Place a mobile conductor carrying a current  $I$  between the poles of a horseshoe magnet as shown in Fig. 25.13. The magnetic force  $F_A$  (termed *Ampere force*) pulls the conductor into the gap between the poles. A reversal of the current causes the conductor to move in the opposite direction.

The direction of the force  $F_A$  acting on a straight current-carrying conductor in a magnetic field is found with the aid of the left-hand rule (Fig. 25.14): if one places one's left hand so that the fingers point in direction of the current in the conductor, with the lines of magnetic induction directed into the palm, the outstretched thumb points in the direction of the force acting on the current-carrying conductor (Ampere force).

Fig. 25.13 Force acts on current-carrying conductor in magnetic field.



\* The term for the unit of measurement of  $\mu_0$  and  $\mu_m$  in the SI system is *henry per metre* (H/m); see Section 26-1.

Ampere demonstrated that  $F_A$  is directly proportional to the length of the conductor  $l$  and the current  $I$  flowing in it. It also depends on the angle  $\alpha$  which the current makes with the lines of induction in the region of the conductor

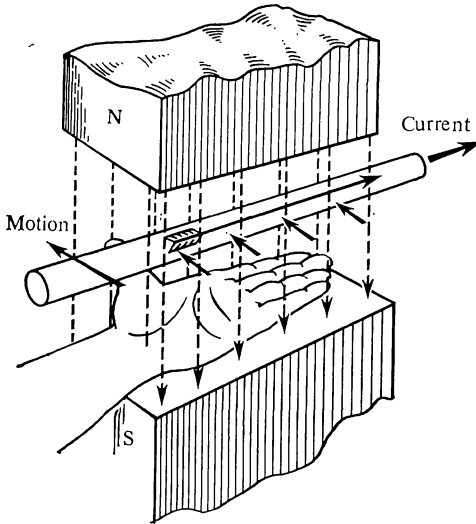


Fig. 25.14 Left-hand rule for determining Ampere force.

(Fig. 25.15). It was established that the force  $F_A$  is proportional to  $\sin \alpha$ , its maximum value,  $F_{A \max}$  being achieved when the conductor is perpendicular to the induction lines. Hence, the formulae for the Ampere force are

$$F_A = BIl \sin \alpha \quad (25.6)$$

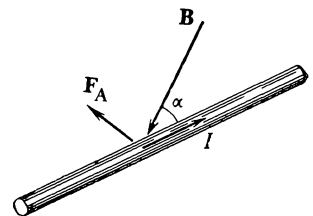
$$F_{A \max} = BIl \quad (25.6a)$$

The factor  $B$  in formulae (25.6) and (25.6a) expresses the dependence of the ampere force on the magnetic field in which the current is placed.

A remarkable feature of the motion of a conductor under the action of an Ampere force is that it is accompanied by the transformation of electric energy into mechanical. This phenomenon is the basic principle of operation of electric motors.

Let us find the physical meaning of factor  $B$  in formula (25.6a). Placing the same conductor carrying current  $I$  in different magnetic fields, one immediately finds that the force  $F_{A \max}$  varies in magnitude and direction. Since  $I$  and  $l$  are in this case constant, it means that  $B$  is variable. It follows from (25.6a) that  $F_{A \max}$  will be greater in a field with greater  $B$ . Similar findings are obtained if the

Fig. 25.15 Ampere force acting on current-carrying conductor in magnetic field depends on angle between current and magnetic induction vector.



conductor is placed in different regions of the same field. Since  $F_{A \text{ max}}$  increases with  $B$ , the factor  $B$  can conveniently be accepted as the measure of the strength of the magnetic field, because in the region of the current-carrying conductor it is the only variable. We obtain from (25.6a)

$$B = \frac{F_{A \text{ max}}}{Il} \quad (25.6b)$$

This formula is valid if the field does not change along the conductor. However, it is generally possible to choose a conductor of such a small length  $\Delta l$  that the changes in the field along it are imperceptible. In such a case  $B$  will be a characteristic of the field at any specific point in it:

$$B = \frac{F_{A \text{ max}}}{I \Delta l} \quad (25.6c)$$

The term for the quantity  $B$ , which is the characteristic of specific points of the field, is *magnetic induction*. The measure for magnetic induction is the force acting on a unit length of a conductor passing through a specific point at right angles to the induction lines and carrying a unit current. It should be pointed out here that magnetic induction is a vector whose direction is determined by the position of a magnetic needle (formula (25.6b) determines only the numerical value of the magnetic induction). Vector  $\mathbf{B}$  is at any point of the magnetic field tangent to the line of magnetic induction passing through this point and points in the same direction as the N pole of a magnetic needle.

We reduce a unit for measuring magnetic induction  $B$ , making use of formula (25.6b):

$$B = \frac{1 \text{ N}}{1 \text{ A} \cdot 1 \text{ m}} = 1 \frac{\text{N}}{\text{A} \cdot \text{m}} = 1 \frac{\text{kg}}{\text{s}^2 \cdot \text{A}} = 1 \text{ T (tesla)}$$

The unit for measuring magnetic induction in the SI system is the *tesla*—magnetic induction of a homogeneous field in which a force of 1 N acts on every metre of length of a conductor carrying a current of 1 A.

By general consent, the number of induction lines crossing a unit area of a surface perpendicular to the lines should be made proportional to  $B$  in the location of the area. This means that the density of induction lines in a diagram is greater in places of greater magnetic induction (sometimes  $\mathbf{B}$  is called *magnetic flux density*).

## 25-10 The Homogeneous Magnetic Field

The parallelism of induction lines is not the only peculiarity of the magnetic field inside a solenoid (see Fig. 25.7). Actually, the magnetic induction vectors at all points of the field are identical in magnitude and direction. The term for such a field is *homogeneous*, or *uniform*. The number of induction lines per unit area of a cross section perpendicular to the induction lines is in this case everywhere the same. Accordingly, the separation of adjacent induction lines should in a homogeneous field be the same everywhere.

A homogeneous magnetic field can be found not only inside a solenoid. The magnetic field in a gap between the opposite poles of a magnet is also homogeneous, provided the gap is much smaller than the dimensions of the poles (Fig. 25.16). Note that at the poles' edges the field can not be regarded as homogeneous.

It has been established by experiment that the only force acting on a closed current-carrying loop or on a magnetic needle is a force couple (Fig. 25.17). In this case magnetic forces can cause only rotational motion. In a nonhomogeneous field a current-carrying loop is forced to move in the direction in which the magnitude of magnetic induction increases.

## 25-11 Magnetic Moment of a Current Loop

It can be proved that the torque  $M$  acting on a loop carrying a current  $I$  in a homogeneous magnetic field is directly proportional to the area of the loop  $A$ , to the current  $I$  and to the magnetic induction  $B$ . Moreover, the torque  $M$  depends on the position of the loop with respect to the field. The torque is at its maximum value  $M_{max}$  when the plane of the loop is parallel to the lines of magnetic induction (see Fig. 25.17), the appropriate expression for it being

$$M_{max} = BIA \quad (25.7)$$

Denoting  $IA = m$ , we obtain

$$M_{max} = mB \quad (25.8)$$

The quantity  $m$ , characterizing the magnetic properties of a current loop which determine its behaviour in an external magnetic field, is called the *magnetic moment* of the loop. The measure for the *magnetic moment* of a loop is the product of the current flowing in it and the area of the loop:

$$m = IA$$

Fig. 25.16 Homogeneous magnetic field between unlike poles of magnet.

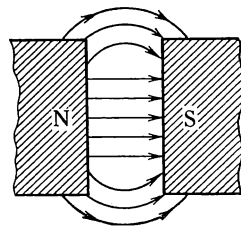
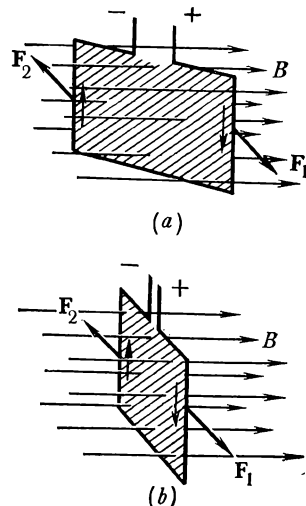
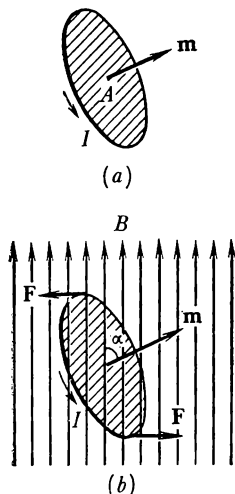


Fig. 25.17 In homogeneous magnetic field, current loop acted upon by forces  $F_1$  and  $F_2$  turns from (a) to (b); directions of  $F_1'$  and  $F_2'$  can be found by left-hand rule.



**Fig. 25.18** (a) Direction of  $\mathbf{m}$  is determined by right-hand screw rule; (b) in homogeneous magnetic field, torque acts on current loop.



(Demonstrate that the unit for measuring  $m$  in the SI system is  $1 \text{ A} \cdot \text{m}^2$ .)

The magnetic moment is a vector whose direction is determined with the aid of the right-hand screw rule: if the screw is rotated in the direction of the current flowing in the loop, the translational motion of the screw will indicate the direction of vector  $\mathbf{m}$  (Fig. 25.18a). The dependence of the torque  $M$  on the orientation of the loop is expressed by the formula

$$M = mB \sin \alpha \quad (25.9)$$

where  $\alpha$  is the angle between the vectors  $\mathbf{m}$  and  $\mathbf{B}$  (Fig. 25.18b).

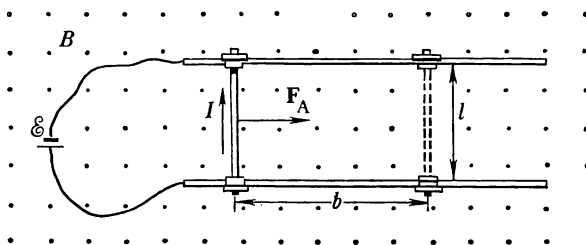
It follows from Fig. 25.18b that a loop can be in a state of equilibrium in a magnetic field when the directions of the vectors  $\mathbf{B}$  and  $\mathbf{m}$  coincide. (When is this equilibrium a stable one?)

## 25-12 Work Done in Moving a Current-Carrying Conductor in a Magnetic Field

Since there is a force acting on a current-carrying conductor in a magnetic field, its displacement will obviously involve work.

Connect two copper rods to a power source (Fig. 25.19) and short-circuit them with a sliding conductor  $l$ . This will cause a current  $I$  to flow in the circuit. If a homogeneous magnetic field of induction  $\mathbf{B}$  is set up in the space surrounding the loop at right angles to it, an Ampere force  $\mathbf{F}_A$  will

**Fig. 25.19** Work of displacing conductor in magnetic field depends on current flowing in conductor and on variation of magnetic flux through loop (dots indicate magnetic field directed towards reader).



act on the conductor  $l$ , causing it to move to the right (prove this). We will now calculate the work performed in displacing the conductor  $l$  by a distance  $b$ .

Since in this case the force and displacement coincide in direction and since  $F_A = BIl$  we have

$$W = F_A b = BIlb$$

If we denote the area of the loop (see Fig. 25.19) with the conductor  $l$  in the initial position by  $A_1$ , and with the conductor  $l$  in its final position by  $A_2$ , then  $\Delta A = A_2 - A_1$  will be the variation in the area of the loop due to the displacement of the conductor  $l$ . It can be seen from Fig. 25.19 that  $\Delta A = lb$ ; therefore

$$W = IB \Delta A$$

Denoting the product  $BA$  by  $\Phi$  (Greek letter *phi*), we obtain

$$B \Delta A = B (A_2 - A_1) = BA_2 - BA_1 = \Phi_2 - \Phi_1 = \Delta \Phi$$

Hence, the work performed in displacing a current-carrying conductor in a magnetic field is expressed by the formula

$$W = I \Delta \Phi \quad (25.10)$$

Let us find out the physical meaning of the quantity  $\Phi$ . Since the value of  $B$  is numerically equal to the number of induction lines passing through unit area  $A_\perp$  perpendicular to them, it follows that  $\Phi = BA_\perp$  is the total number of lines of magnetic induction crossing the area  $A_\perp$  (if magnetic induction  $B$  is the same at all points in  $A_\perp$ ). The accepted term for  $\Phi$  is *magnetic flux*, or the flux of vector  $\mathbf{B}$  through  $A_\perp$ .

Hence, in a homogeneous field the measure of the magnetic flux is the product of  $B$  and  $A_\perp$ :

$$\Phi = BA_\perp \quad (25.11)$$

Note that the magnetic flux is a scalar.

We deduce a unit for measuring the magnetic flux:

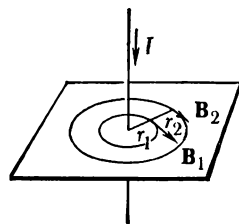
$$\Phi = 1 \text{ T} \cdot 1 \text{ m}^2 = 1 \text{ T} \cdot \text{m}^2 = 1 \frac{\text{kg} \cdot \text{m}^2}{\text{s}^2 \cdot \text{A}} = 1 \text{ Wb (weber)}$$

The unit of  $\Phi$  in the SI system is the *weber*, which is the magnetic flux through an area of  $1 \text{ m}^2$  perpendicular to the induction lines in a magnetic field with  $1 \text{ T}$  induction.

### 25-13 Magnetic Induction Due to Currents in Conductors of Different Shape

The French physicists Jean Baptiste Biot (1774-1862) and Felix Savart (1791-1841) demonstrated in 1820 that the induction of the magnetic field of a straight current at a point in space is directly proportional to the current,  $I$ , and inversely proportional to the distance of the point from the conductor,  $r$  (Fig. 25.20).

Fig. 25.20 Induction  $\mathbf{B}$  of magnetic field of linear current is inversely proportional to distance  $r$  from conductor with current.





Indeed, we know already that a force  $F_{A \max} = BIl$  acts on a conductor of length  $l$  carrying a current  $I$  in a magnetic field with induction  $B$ . Hence

$$B = \frac{F_{A \max}}{Il}$$

Let  $F_{A \max}$  be the force with which the field of the current  $I_1$  acts on a section of length  $l$  of another parallel conductor carrying a current  $I_2$  (Fig. 25.12). In this case  $B$  indicates the induction  $B_1$  of the first current  $I_1$  at the location of the second current  $I_2$ . Using the relation (25.4) for  $F_{A \max}$ , we obtain

$$B_1 = \frac{F_{A \max}}{I_2 l} = \frac{\mu_m I_1 I_2 l}{2\pi a I_2 l} = \mu_m \frac{I_1}{2\pi a}$$

Substituting  $r$  for  $a$  and omitting the indices, we obtain the formula for calculating the magnetic induction of a straight current:

$$B_s = \mu_m \frac{I}{2\pi r} \quad (25.12)$$

The induction of a magnetic field in a current-carrying conductor of arbitrary shape at each point of space is determined by the vector sum of the magnetic inductions set up by individual sections of the conductor and can be calculated theoretically. Here are the appropriate formulae for two important examples.

The induction at the centre of a circular current is expressed by the formula

$$B_c = \mu_m \frac{I}{2r} \quad (25.13)$$

where  $r$  is the radius of the circular current.

The induction of a magnetic field inside a solenoid of length  $l$  greatly exceeding its diameter, of  $w$  turns and carrying a current  $I$  is expressed by the formula

$$B_{sol} = \mu_m \frac{Iw}{l} \quad (25.14)$$

Since the magnetic field in such a solenoid is homogeneous, the magnetic flux can be expressed by formula (25.14):

$$\Phi_{sol} = B_{sol} A$$

where  $A$  is the cross section of the solenoid. Substituting the relation (25.14) for  $B_{sol}$ , we obtain the formula for calculating the magnetic flux of a solenoid

$$\Phi_{sol} = \mu_m \frac{Iw}{l} A \quad (25.15)$$

The usual term for the product  $Iw$  is the *ampereturns* of the solenoid, sometimes called its *magnetomotive force*.

### 25-14 Magnetic Field Strength

It has been established by experiment that the permeability  $\mu_m$  of the great majority of homogeneous substances remains constant over a wide range of induction  $B$ . This enables a new physical quantity  $H$  to be introduced, to which the magnetic induction  $B$  in a medium of constant permeability  $\mu_m$  should be proportional:

$$B = \mu_m H \quad (25.16)$$

But why should we introduce a new quantity,  $H$ , if we already have  $B$ ? The answer is as follows. If we travel along a circular path, so that we finally return to the place where we started, and on each small segment of that path calculate the product of  $B$  and the length of the segment, and then add all the products, we find a quantity known as the *circulation* of  $B$  over the closed path. It so happens that the result is determined by all the currents inside the "loop": both macrocurrents (which we can measure by an ammeter) and microcurrents (molecular currents, for example; see Section 25-3). But this is inconvenient, since we know how to measure macrocurrents directly, but do not know how to measure microcurrents. So we would like a quantity whose circulation along a closed path is determined only by macrocurrents, which is exactly what  $H$  is. This does not mean, however, that  $H$  is determined solely by macrocurrents. (Only its circulation is determined by macrocurrents.) The basic quantity is  $B$ , and not  $H$ . But it is sometimes more convenient to work with  $H$  than  $B$ . So the role of  $\mu_m$  in (25.16) is to "separate" the medium from the field. [For more detail see E. M. Purcell: *Electricity and Magnetism* (Berkeley Physics Course, vol. 2), McGraw-Hill, New York, 1965, Section 10-10.] Finally, we note that  $H$  is called the *magnetic field strength* (formerly magnetic intensity).

Note that the magnetic field strength is a vector, whose direction in an isotropic medium coincides with vector  $\mathbf{B}$ .

Comparing formulae (25.12) and (25.16) we obtain the formula for the magnetic field strength of a straight current:

$$H = \frac{I}{2\pi r} \quad (25.17)$$

(Demonstrate that the magnetic field strength in the centre of a circular current

$$H = \frac{I}{2r} \quad (25.18)$$

and at the centre of a solenoid

$$H = \frac{Iw}{l} . ) \quad (25.19)$$

The unit for measuring  $H$  can be deduced from formula (25.17):

$$H = \frac{1 \text{ A}}{2\pi (1/2\pi) \text{ m}} = 1 \frac{\text{A}}{\text{m}}$$

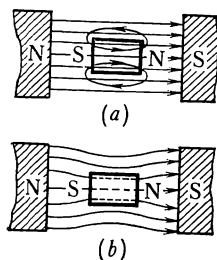
The unit for measuring magnetic field strength in the SI system is the strength of a magnetic field set up by a current of 1 A flowing in a long straight conductor  $1/2\pi$  m from it.

The unit for measuring  $\mu_m$  can be deduced from formula (25.16):

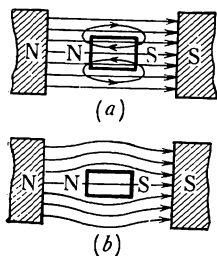
$$\mu_m = \frac{B}{H}, \quad \mu_m = \frac{1 \text{ T}}{1 \text{ A/m}} = 1 \frac{\text{T} \cdot \text{m}}{\text{A}}$$

The unit for measuring the magnetic permeability  $\mu_m$  in the SI system is the magnetic permeability of a medium in which a magnetic field strength of 1 A/m sets up a magnetic induction of 1 T (see also Sections 25-8 and 26-1).

**Fig. 25.21** (a) Superposition of magnetic field of paramagnetic substance on external magnetic field; (b) magnetic field resulting from superposition of field of magnetized paramagnetic substance on external field.



**Fig. 25.22** (a) Superposition of magnetic field of diamagnetic substance on external field; (b) magnetic field resulting from superposition of field of diamagnetic substance on external field.



### 25-15 Paramagnetic, Diamagnetic and Ferromagnetic Substances

A substance that can be magnetized in a magnetic field is termed *magnetic substance*. Some substances magnetized by a field augment it, others weaken it. Consider first substances whose molecules possess intrinsic magnetic fields set up by electrons orbiting the nuclei. Such a magnetic field is similar to the magnetic field of a circular current, and therefore such molecules can be seen as tiny magnets, each having an N pole and an S pole.

If such a substance is placed in a magnetic field, torques act on its molecules, arranging them in a regular order along lines of magnetic induction so that the induction lines enter a molecule from its S pole and leave it from its N pole. Hence, the magnetic field is augmented in the substance. Bodies made of such substances are magnetized by external fields as shown in Fig. 25.21a. The superposition of the field set up by the substance and of the external field produces a resultant magnetic field shown in Fig. 25.21b, in which the induction lines are seen to be drawn into the body. A rod made of such a substance arranges itself in the direction of the induction lines.

Since the thermal motion of the molecules disturbs their regular order, magnetization decreases with temperature. When such a body is withdrawn from the magnetic field the random motion of its molecules soon causes it to be completely demagnetized.

It follows from the above that the relative magnetic permeability of such a magnetic should exceed unity. Substances with permeabilities a little above  $\mu_0$  are said to be *paramagnetic*. Note that the magnetization of paramagnetic substances is quite weak even in very strong external fields. It follows from Fig. 25.21 that a paramagnetic substance tends to be drawn into the external magnetic field, since unlike magnetic poles attract each other. The explanation for paramagnetism of substances is the orbital motion of electrons about atomic nuclei, a motion which produces intrinsic molecular magnetic fields.

The behaviour in an external magnetic field of substances whose molecules have no intrinsic magnetic field is different. A body made of such a substance is magnetized in such a way that the intrinsic magnetic field inside the body is directed against the external field (Fig. 25.22a). The relative permeability of such substances is less than unity.

Substances with a magnetic permeability of a little less than  $\mu_0$  are said to be *diamagnetic*. Diamagnetic properties are much less manifest than paramagnetic. Bismuth is a typical diamagnetic element. It follows from 25.22b that a diamagnetic substance tends to be expelled from an external magnetic field because like magnetic poles repel each other. The origin of diamagnetism will be discussed in the following chapter (see Section 26-5). Numerical values for the relative permeability of some substances are presented in Table 25.1.

**Table 25.1** Relative permeability of some substances

Diamagnetic	$\mu$	Paramagnetic	$\mu$
Air (gas)	1.00000038	Antimony	0.999937
Aluminium	1.000023	Benzene	0.999993
Ebonite	1.000014	Bismuth	0.999824
Iridium	1.000063	Copper	0.999991
Manganese	1.0038	Germanium	0.9999986
Nitrogen (gas)	1.000000013	Glass	0.999987
Oxygen (gas)	1.0000019	Gold	0.999963
Oxygen (liquid)	1.0034	Hydrogen (gas)	0.999999937
Palladium	1.000692	Lead	0.999987
Platinum	1.000360	Quartz	0.999985
Tin	1.000044	Rock salt	0.999987
Wolfram	1.000175	Silicon	0.999837
		Silver	0.999981
		Water	0.999991
		Zinc	0.999988

Fig. 25.23 Magnetic field resulting from superposition of field of ferromagnetic substance on external field.

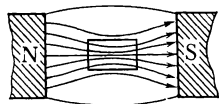
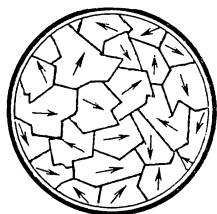
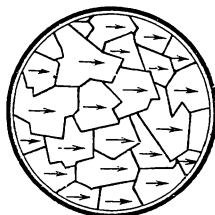


Fig. 25.24 Arrangement of magnetic domains in (a) steel not yet magnetized and (b) steel already magnetized.



(a)



(b)

Besides the substances described above there is a small group of substances with a relative permeability greatly exceeding unity. Substances whose permeability greatly exceeds  $\mu_0$  are said to be *ferromagnetic*. The most striking example of such substances is iron. In some cases it can amplify an external magnetic field by a factor of several thousand. Other ferromagnetics include steel, pig iron, nickel, cobalt, the rare metal gadolinium and some alloys of ferromagnetic metals. The “pulling in” of lines of magnetic induction of an external field is very pronounced in a ferromagnetic substance (Fig. 25.23).

Microscopic studies of ferromagnetic substances helped establish the fact that they consist of numerous regions about 0.001 mm in size that can be spontaneously magnetized. These regions are termed *domains*. All molecular magnetic moments inside a domain coincide in direction.

In a nonmagnetized ferromagnetic substance the domains are arranged at random (Fig. 25.24a). When a ferromagnetic substance is placed in an external magnetic field, the orientation of the internal field of the domains changes so as to make their magnetic moments coincide in direction with the induction lines of the external field thus greatly augmenting it (Fig. 25.24b).

Only substances with these domains are ferromagnetic. The maximum magnetization of a ferromagnetic substance occurs when all its domains are oriented in the direction of the external field. Such a state of a ferromagnetic substance is termed *magnetic saturation*. Note that each individual domain is always in a state of magnetic saturation.

The explanation for ferromagnetism was found after it had been established that in addition to its orbital motion an electron also rotates about its own axis, that is, possesses an intrinsic angular momentum subsequently termed *spin*. But this implies that the electron also has an intrinsic magnetic moment. The magnetic moments of all electrons in an atom can only be oriented either parallel or antiparallel to each other. The majority of electrons in all atoms form pairs in which the magnetic moment of one electron is opposed to that of the other. Because of this the electron magnetic moments in nearly all atoms are compensated.

The atoms of the ferromagnetic elements each have several electrons with uncompensated magnetic moments (i.e. pointing in one direction). These electrons interact with neighbouring atoms in a way somewhat similar to that of electrons in a covalent bond. This turns the magnetic moments of all electrons in a domain in one direction.

Hence, the magnetic properties of ferromagnetic substances are essentially due to the electrons of their atoms possessing uncompensated magnetic moments and to exchange coupling of these electrons.

### 25-16 Magnetization of Ferromagnetic Substances

The magnetic induction in magnetized para- and diamagnetic substances varies in direct proportion to the magnetic field strength (Fig. 25.25).

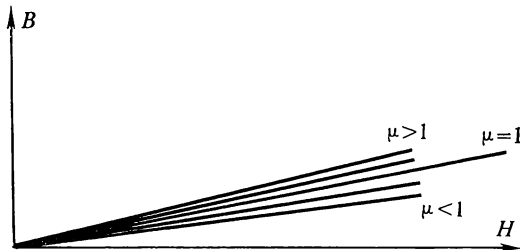


Fig. 25.25 Magnetic induction versus magnetic field intensity for paramagnetic ( $\mu > 1$ ) and diamagnetic ( $\mu < 1$ ) substances (departure from straight line  $B = \mu_0 H$  for vacuum ( $\mu = 1$ ) is greatly exaggerated).

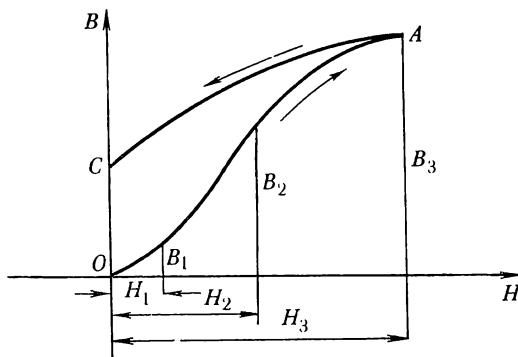


Fig. 25.26 Magnetic induction versus magnetic field intensity for steel.

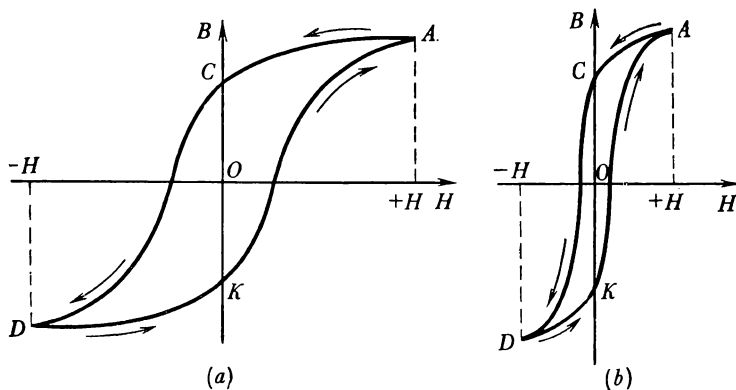
The magnetization of a ferromagnetic substance is different (Fig. 25.26). At first an increase in  $H$  is accompanied by a very fast rise in inductance. Its rise then slows down, and at large enough values of  $H$  we find that  $B$  practically ceases to increase. If the experiment is made with a non-magnetized ferromagnetic substance, the magnetization proceeds along the curve  $OA$ , the term for which is the *virgin curve* of magnetization. It can be seen from the graph that  $\mu_m = B/H$  for a small  $H$  is not great; then it increases; and finally begins to diminish. This means that the perme-

ability of ferromagnetic substances does not remain constant with  $H$ .

The explanation for the shape of the curve shown in the figure is as follows. As long as the orientation of the domains in the direction of the external field occurs, induction rises swiftly. When the substance becomes magnetized to saturation point, the subsequent rise in  $B$  will be entirely at the expense of the rise in  $H$ . If one starts now to gradually reduce the field strength, the demagnetization proceeds along the curve  $AC$ . Note that at  $H = 0$  the ferromagnetic substance remains magnetized, since here its induction corresponds to the segment  $OC$ . Hence, the induction  $B$  in a ferromagnetic substance depends not only on  $H$  but also on the state of its initial magnetization.

It can be seen from Fig. 25.26 that in the process of demagnetization induction follows a curve of lesser slope

Fig. 25.27 Hysteresis loops for (a) hard magnetic steel; (b) soft magnetic steel.



than the magnetization curve. The term for this phenomenon is *magnetic hysteresis* (lag). When the ferromagnetic substance is periodically remagnetized with an alternating magnetic field, the  $B$ - $H$  curve forms the so-called *hysteresis loop* (Fig. 25.27). The area of the hysteresis loop has been found to be proportional to the energy spent on remagnetizing the dielectric. This energy turns into the internal energy of the ferromagnetic substance. Consequently, periodic remagnetization should raise the temperature of the ferromagnetic substance. Ferromagnetic substances with a hysteresis loop of large area are termed *hard magnetic materials* (they are used for fabricating permanent magnets), the term for ferromagnetic substances with a small hysteresis loop being *soft magnetic materials* (Fig. 25.27). In com-

paratively recent times materials with hysteresis loops of extra-small area, termed *ferrites*, have been produced. Their use makes it possible to reduce remagnetization losses.

It has been established by experiment that the magnetic properties of ferromagnetic substances are temperature-dependent. Heating results in a decrease in the permeability of ferromagnetic substances. If the temperature is high enough, their domains break up and they become paramagnetic. The temperature at which such a transformation takes place is called the *Curie point* (the Curie point of iron is  $770^{\circ}\text{C}$  and that of nickel is  $360^{\circ}\text{C}$ ). Upon cooling the substance again becomes a ferromagnetic.

The "pulling in" of induction lines by a ferromagnetic substance is used for *magnetic shielding*. If a casing is made from a ferromagnetic substance, the induction lines of an external field will pass along its walls and so there will be no field inside the casing (Fig. 25.28). This method is used to protect sensitive instruments from the magnetic fields of the Earth and other magnetized bodies.

The amplification of a magnetic field by ferromagnetic substances is widely used in engineering practice. For example, the amplification of the magnetic field of a solenoid by a ferromagnetic substance is widely used in electric magnets. The rod inserted into the solenoid is termed the *core*. A solenoid with a core made from soft magnetic steel is called an *electromagnet*, while the wire with which the solenoid is made is called the *winding of the electromagnet*. Often an electromagnet is made in the shape of horseshoe. A schematic representation of such an electromagnet is shown in Fig. 25.29.

A remarkable property of the electromagnet is that it can be magnetized or demagnetized by turning the current in its winding on or off. It is this property that gives it such wide application in various automatic devices, such as the electromagnetic relay. In modern electrical engineering electromagnets are used in cranes, telephones, telegraphy, electric motors, generators, test instruments, etc.

One of the important properties of ferromagnetic substances is the variation of their volume in the process of remagnetization. The term for this property is *magnetostriction*. It is used for the generation of ultrasonic vibrations (see Section 28-8). To this end a core with a protruding end is inserted into a coil and a current of the appropriate ultrasonic frequency is made to pass through the coil.

Fig. 25.28 Magnetic shielding.

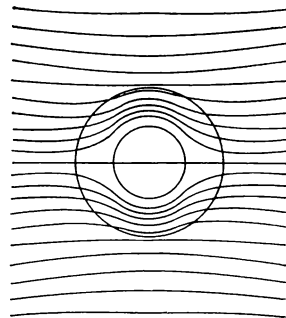
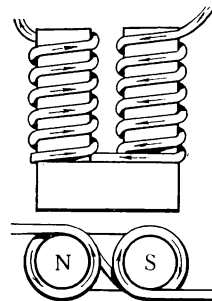
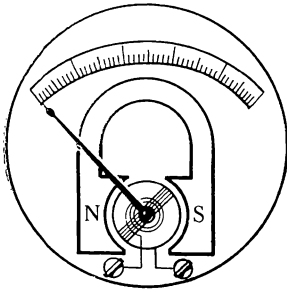


Fig. 25.29 Electromagnet (electromagnet's poles shown below).

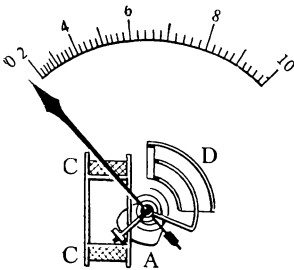




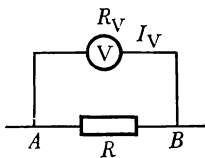
**Fig. 25.30** Moving-coil instrument.



**Fig. 25.31** Moving-iron instrument.



**Fig. 25.32** Measuring voltage between points A and B with voltmeter.



## 25-17 Construction of an Ammeter and a Voltmeter

The operation of most electric measuring instruments is based on the magnetic effect of an electric current. Consider the design of two types of such instruments: the moving-coil and the moving-iron types.

In a *moving-coil instrument* there is a permanent magnet which does not move and a coil which is turned by the Ampere force when there is a current flowing in it (Fig. 25.30). A spiral spring on the coil's axis opposes the rotation of the coil. The greater the current flowing in the coil the greater its deflection angle. The coil is joined to a pointer whose point moves across a scale. Moving-coil instruments have great accuracy and high sensitivity, but can be used only with direct current.

A *moving-iron instrument* (Fig. 25.31) has a static coil C and a moving core A, made of soft magnetic steel, which is drawn into the coil when current flows in it. The core is joined to a pointer whose point moves across a scale. The vibrations of the pointer after the connection of the instrument into a circuit are arrested by an air brake D, called a *damper*. This instrument is not as accurate or sensitive as the *moving-coil* type but it can be used both in dc and ac circuits and can withstand overloads.

The connection of a measuring instrument into a circuit must not introduce perceptible changes into the operation of the electric circuit. For instance, the connection of an ammeter or a voltmeter must not cause the current in the circuit to change.

Note that the design of the ammeter is identical to that of the voltmeter, except for its internal resistance. An ammeter is connected into a circuit in series and because of that its resistance should be as small as possible; otherwise its connection will result in a noticeable decrease in the current. A voltmeter is connected in parallel to the two points the voltage between which it measures and therefore its resistance should be as great as possible.

The voltage between points A and B (Fig. 25.32) is equal to the product  $IR$  in one of the branches. If this branch is a voltmeter, then  $U_{AB} = I_V R_V$ . Since  $R$  is constant the voltage  $U_{AB}$  is proportional to the current  $I_V$  flowing in the voltmeter. This means that a voltmeter is, in fact, an ammeter whose scale is graduated according to the values of the product of the current flowing in the instrument  $I_V$  and its resistance  $R_V$ .

If an ammeter designed to measure currents up to  $I_A$  is then required to measure currents up to  $nI_A$ , a *shunt*

(Fig. 25.33a) is connected in parallel to it with a resistance several times less than that of the ammeter,  $R_A$ . It follows from Fig. 25.33a that

$$I_{\text{sh}}/I_A = R_A/R_{\text{sh}}$$

Since  $I_{\text{sh}} = I - I_A$ , we obtain

$$I/I_A - 1 = R_A/R_{\text{sh}}$$

Since it is specified that  $I/I_A = n$ , we have  $n - 1 = R_A/R_{\text{sh}}$ , whence

$$R_{\text{sh}} = R_A/(n - 1) \quad (25.20)$$

Formula (25.20) enables us to calculate the resistance of a shunt designed to extend the range of an ammeter  $n$  times.

To be able to use a voltmeter designed to measure voltages not exceeding  $U_V$  for measuring voltages  $n$  times higher, a high-resistance coil of wire, or *multiplier*,  $R_M$ , is connected in series with the voltmeter (Fig. 25.33b). The maximum voltage which the voltmeter is now able to measure is equal to the sum  $U_V + U_M$ . Since  $U_M/U_V = R_M/R_V$  and  $U_M = U - U_V$ , we have

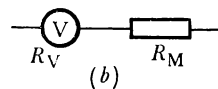
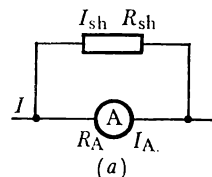
$$U/U_V - 1 = R_M/R_V, \quad \text{or} \quad n - 1 = R_M/R_V$$

whence

$$R_M = R_V (n - 1) \quad (25.21)$$

Formula (25.21) enables us to calculate the value of a multiplier for a voltmeter designed to extend its range  $n$  times.

Fig. 25.33 Connecting shunt to ammeter (a) and multiplier to voltmeter (b).



## 25-18 The Lorentz Force Equation

The Dutch physicist Hendrick Antoon Lorentz (1853-1928) identified an Ampere force with the influence of a magnetic field on charges moving in a current-carrying conductor. Since these charges are unable to leave the conductor, the total force acting on them is transmitted to the conductor.

Thus, the ampere force  $F_A$  is the sum of the forces acting on the free charges in a current-carrying conductor. This hypothesis enables the force acting on an individual mobile charge in a magnetic field to be calculated. The accepted term for this force is the *Lorentz force*. Hence

$$F_{\text{Lor}} = F_A/N$$

where  $N$  is the total number of free charges in the conductor. In a metal such charges are the electrons, each with a charge

Fig. 25.34 Charge  $q$  in homogeneous magnetic field moves in circle of radius  $r$  (dots indicate magnetic field directed towards reader).

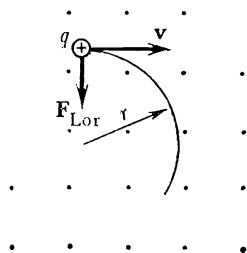
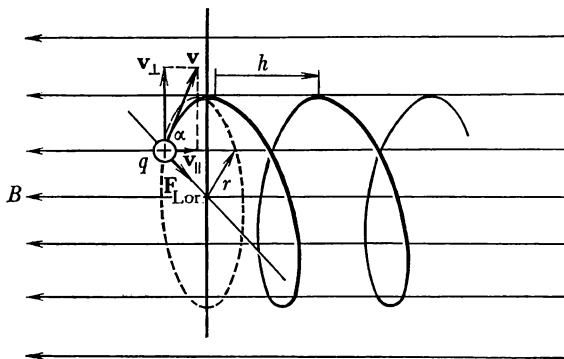


Fig. 25.35 When velocity  $\mathbf{v}$  of charge  $q$  makes angle  $\alpha$  with induction lines,  $q$  "winds" itself around induction lines.



$e$ . Since  $F_A = BIl \sin \alpha$  and  $I = vn_0eA$  (see Section 18-2), it follows that

$$F_{\text{Lor}} = \frac{RIl}{N} \sin \alpha = \frac{Bvn_0eAl}{N} \sin \alpha = \frac{Bvn_0eV}{N} \sin \alpha$$

Taking into account that  $n_0V = N$ , we obtain a formula for calculating the Lorentz force

$$F_{\text{Lor}} = Bve \sin \alpha \quad (25.22)$$

where  $\alpha$  is the angle between the vectors  $\mathbf{B}$  and  $\mathbf{v}$ .

The direction of the Lorentz force is found with the aid of the *left-hand rule* (see Section 25-9). When applying it, we should remember that if a positive charge  $e_+$  moves in a magnetic field then the four outstretched fingers should point in the direction of its motion, that is, in the direction of vector  $\mathbf{v}$ , and if the charge is negative the outstretched fingers should point in the direction opposite to  $\mathbf{v}$ .

The Lorentz force acts always at right angles to the plane containing vectors  $\mathbf{B}$  and  $\mathbf{v}$ . This means that it is perpendicular to both these vectors and therefore cannot perform work, that is, cannot change the kinetic energy of free charges moving in a magnetic field. It can only change the direction of the velocity of motion of the free charges, this is, it is a centripetal force.

Suppose a charge  $q$  with a mass  $m$  and a velocity  $\mathbf{v}$  flies into a homogeneous magnetic field with an induction  $\mathbf{B}$  so that the vector  $\mathbf{v}$  is perpendicular to  $\mathbf{B}$ . Then  $F_{\text{Lor}} = F_c$ :

$$qBv = mv^2/r$$

In this case the charge will move in a circle (Fig. 25.34) with radius

$$r = (m/qB) v \quad (25.23)$$

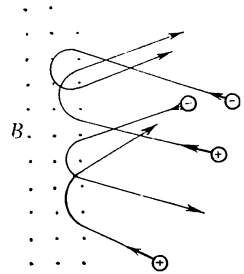
If the direction of velocity makes an angle  $\alpha$  other than  $90^\circ$  with the induction lines, the charge will move in a helix around the induction lines of the field (Fig. 25.35), since vector  $\mathbf{v}$  can be resolved into two components  $v_{\parallel}$  and  $v_{\perp}$ , one of which,  $v_{\parallel}$  coincides in direction with the induction lines and the other,  $v_{\perp}$ , is normal to them. The latter determines the radius of the spiral, and  $v_{\parallel}$  remaining unaffected. If the charge completes a turn in time  $T$ , its displacement in the direction of the induction line during this time will be  $h = v_{\parallel}T$ . It can be easily seen that  $v_{\parallel} = v \cos \alpha$  and  $v_{\perp} = v \sin \alpha$  and  $h$  is the pitch of the helix.

For a charged particle moving in a nonhomogeneous field not only the direction, but the magnitude of the Lorentz force will be variable and the particle's path may be very intricate.

Consider now the case of a charged particle entering a strong magnetic field, as shown in Fig. 25.36. If the particle moves in a plane perpendicular to the induction lines, on entering the magnetic field it will travel along an arc (whose radius is determined by formula (25.23)) and then fly out of the field. If the particle flies in at an arbitrary angle to the induction lines, it will still, after covering some part of the turn of the spiral, fly out of the field.

Hence, strong magnetic fields reflect charged particles entering them and for this reason are sometimes termed *magnetic mirrors*. Note that this property of a magnetic field is utilized in nuclear physics for confining high-temperature plasma. Strong magnetic fields are set up around the plasma; the fields reflect the plasma's charged particles and thus play the part of a vessel capable of containing the plasma.

**Fig. 25.36** Strong magnetic field rejects charged particles trying to enter it (induction lines directed at reader).



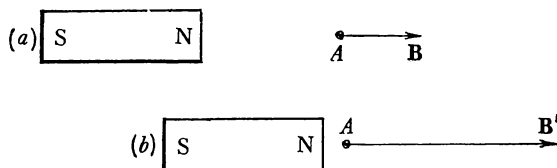
## 25-19 Constant and Variable Magnetic Fields

Let us examine the conditions in which a magnetic field in space is either constant or variable. Suppose a magnetic field is set up by a permanent magnet and that the observer is at point  $A$  (Fig. 25.37), where the induction of the field is  $\mathbf{B}$ . As long as the magnet remains at rest the induction  $\mathbf{B}$  at point  $A$  will remain constant. If the magnet is set in motion,  $\mathbf{B}$  will begin to change. For instance, if the magnet is brought closer to point  $A$ , the induction  $\mathbf{B}$  at this point will rise to some new value  $\mathbf{B}'$ . It is obvious that the change in induction  $\mathbf{B}$  at point  $A$  is the result of the motion of the magnet and that after the magnet stops the induction  $\mathbf{B}'$  will no longer change.

The term *constant magnetic field* applies to a field at every point of which the vectors of magnetic induction remain constant. A constant magnetic field exists around a permanent magnet or a static conductor carrying a constant direct current.

A *variable magnetic field* is the result not only of the motion of a magnet or a current-carrying conductor with respect to the observer. We recall that the induction of a magnetic field depends on the current flowing in the conductor which produces the field. Accordingly, the magnetic

**Fig. 25.37** As magnet moves from (a) to (b), induction vector grows from  $\mathbf{B}$  to  $\mathbf{B}'$ .



field in the space surrounding a static conductor carrying a variable current is also variable. For instance, when a circuit is closed, the current increases during a short time interval from zero to its stationary value, the magnetic field varying with the current. Conversely, when the circuit is disconnected, the current and its magnetic field drop to zero.

The magnetic field around a conductor carrying an alternating current is also alternating. Note that in this case both the magnitude and the direction of vector  $\mathbf{B}$  are subject to variation.

## 25-20 Magnetic Fields in Solar and Cosmic Phenomena

Studies of the Sun lead to the conclusion that it possesses a magnetic field with a strength about twice that of the Earth's. Many phenomena taking place in the Sun's atmosphere (the formation of dark spots, flares, etc.) are closely related to the appearance and evolution of strong local magnetic fields. Such localities have been termed *active*.

It was mentioned above that an intense mixing of gas, convection, takes place in the layer lying below the photosphere. Investigations produced the result that in the vicinity of a *dark spot* there is always a strong magnetic field with a strength a thousand times greater than the intensity

in unperturbed areas. This field deflects charged particles of plasma and obstructs the formation of convection fluxes. In this region hot gases from underlying layers are no longer able to rise and so the gas here cools down substantially.

In the region of a *flare* the magnetic field is by no means strong enough to stop vertical convection fluxes on plasma. However, it suppresses random motion of the plasma in the flux and thereby reduces internal friction, thus creating the conditions for a stable rising flux of hot gas, the flare.

Many phenomena observed in the atmosphere of the Sun are connected with variable magnetic fields. It was mentioned above that a magnetic field changes only the direction of velocity of a charged particle moving in it. It was established that a variable magnetic field penetrating the plasma can cause changes not only in the direction but also in the magnitude of the velocity of charged particles, and can thus create directional motion of the plasma. This is the mechanism which occasionally results in the formation of powerful plasma fluxes ejecting great masses of gas into the corona to form *prominences*—gigantic gas clouds extending far into the corona (see Fig. 8.3).

A strong magnetic field, changing in the process of the evolution of a group of spots, exerts pressure on the plasma. An intense compression of the plasma in the chromosphere above the region of spot concentration sometimes takes place, causing a substantial rise in the temperature of the gas. A sudden drastic intensification of gas luminescence is observed in this part of the chromosphere termed *chromospheric flare*.

The variable magnetic field ejects into outer space fluxes of plasma particles, called *corpuscular streams*, moving at speeds of about 1000 km/s. Some particles are accelerated to enormous speeds (comparable to the speed of light) to become *solar cosmic rays*.

Observations carried out over the years produced the result that the number and total area occupied by spots change periodically and reach maximum values on the average every eleven years. In these years the number of flares and prominences increases, *chromospheric flares* are observed more frequently, and the intensity of corpuscular radiation rises by a factor of several tens. The general term for all these phenomena is *solar activity*.

On reaching the Earth the fluxes of charged particles ejected from the Sun are deflected by its magnetic field and, in turn, affect it. Great perturbation of the Earth's magnetic field, *magnetic storms*, have been observed to take place at times of maximum solar activity. During such storms

the magnetic needle of a compass wanders at random. Some of the charged particles penetrate the Earth's magnetic field and, moving in spirals along induction lines, find themselves in a sort of a trap. Concentrating in ring-shaped zones around the Earth the charged particles form *radiation belts* detected with the aid of satellites. At the poles the charged particles easily penetrate the atmosphere causing the phenomenon of the *aurora polaris*.

Magnetic fields exist in interstellar space as well. Their intensity is tens of thousands times less than that of the Earth, but being colossal in extent they greatly affect the nature of motion of charged particles in interstellar space.

## 26 Electromagnetic Induction

### 26-1 Flux Linkage and Inductance

In Section 25-12 we gave formula (25.10), which can be used to calculate the work done in rotating a current loop in an external magnetic field. Let us find now the work needed to rotate a solenoid of  $w$  turns in an external magnetic field.

Since the work done in rotating one turn is  $I\Delta\Phi$  and since in this case the magnetic flux penetrates  $w$  turns, the expression for the work will be

$$W = Iw\Delta\Phi = Iw(\Phi_2 - \Phi_1) = I(w\Phi_2 - w\Phi_1)$$

Denoting the product  $w\Phi$  by  $\psi$  (Greek letter *psi*) we obtain the formula for work

$$W = I(\psi_2 - \psi_1), \quad \text{or} \quad W = I\Delta\psi \quad (26.1)$$

The quantity  $\psi$  characterizing the coupling (the linkage) of the magnetic flux with a closed circuit penetrated by it is called the *flux linkage*. For a magnetic flux  $\Phi$  penetrating a coil of  $w$  turns the flux linkage is equal to the product of the number of turns and the magnetic flux:

$$\psi = w\Phi \quad (26.2)$$

(Demonstrate that the unit for measuring flux linkage in the SI system is the weber; see Section 25-12.)

Now let us imagine an arbitrary closed circuit carrying a current  $I$ . This current sets up its own magnetic field around the circuit. Naturally, there is an intrinsic flux  $\Phi$

passing through the surface bounded by the conductors of the closed circuit. If this circuit consists of a single plane loop, then  $\psi = \Phi$ . If the conductors combine to form a coil of  $w$  turns,  $\psi = w\Phi$ . Hence, the intrinsic flux linkage of a circuit depends on its configuration, that is, on the position of the conductors in space.

It is an experimental fact that in the absence of ferromagnetic substances the intrinsic flux linkage of a circuit is proportional to the current flowing in it:

$$\psi = LI \quad (26.3)$$

The proportionality factor  $L$  remains constant only if the configuration of the conductors making up the closed circuit and the medium remain constant. This factor characterizing the dependence of the intrinsic flux linkage of a closed circuit on its shape and on the surrounding medium is called the *inductance* of the circuit. A unit for measuring  $L$  is

$$L = \psi/I, \quad L = 1 \text{ Wb/1 A} = 1 \text{ Wb/A} = 1 \Omega \cdot \text{s} \\ = 1 \text{ H (henry)}$$

The unit for measuring inductance in the SI system is the *henry*. Henry is the term for the inductance of a circuit with a flux linkage of 1 Wb at a current of 1 A.

We recall that the unit for measuring magnetic permeability in the SI system is  $\text{N/A}^2$  (see Section 25-8) or  $\text{T} \cdot \text{m/A}$  (see Section 25-14). Since  $1 \text{ Wb} = 1 \text{ T} \cdot \text{m}^2$  (see Section 25-12), it follows that  $1 \text{ N/A}^2 = 1 \text{ T} \cdot \text{m/A} = 1 \text{ H/m}$ . The commonly used unit is the latter,  $1 \text{ H/m}$ .

By way of an example, let us find the inductance of a solenoid,  $L_{\text{sol}}$ . We have from (26.3)

$$L_{\text{sol}} = \psi_{\text{sol}}/I_{\text{sol}} = w\Phi_{\text{sol}}/I_{\text{sol}}$$

Since  $\Phi_{\text{sol}}$  is found from relation (25.15), it follows that

$$L_{\text{sol}} = \frac{w\mu_m I_{\text{sol}} wA}{l I_{\text{sol}}} = \mu_m \frac{w^2 A}{l} \quad (26.4)$$

Hence, the inductance of a solenoid is determined by the medium, by the number of turns and by the solenoid's dimensions.

## 26-2 Discovery of Induced Current

In Chapter 25 it was established that an electric current and its magnetic field always exist together. Faraday, being aware of the close interrelation of current and magnetic



field, was certain that a current could be produced in a closed circuit with the aid of a magnetic field. He performed numerous experiments and succeeded in proving it after he discovered in 1831 the phenomenon of electromagnetic induction.

When an electric current in a conductor is produced by a change in a magnetic field we speak of *electromagnetic induction*. The current obtained in this way is termed *induced current* and the emf responsible for it *induced emf*.

Comprehensive research into the phenomenon of electromagnetic induction led to the conclusion that it can be used to produce currents of practically unlimited power and this makes possible the extensive supply of electrical energy for industrial needs. At present almost all electric energy is produced by *induction generators*, whose operation is based on electromagnetic induction. This is why Faraday is justly regarded as one of the founders of electrical engineering.

Let us discuss the essence of electromagnetic induction in more detail.

### 26-3 Induced EMF in a Straight Conductor Moving in a Magnetic Field

Suppose a straight metallic conductor of length  $l$  is in a homogeneous magnetic field with an induction  $\mathbf{B}$  (Fig. 26.1). If this conductor is set in motion with a velocity  $\mathbf{v}$  at right angles to  $\mathbf{B}$  ( $\alpha = 90^\circ$ ), its electrons will move with it. Since the electrons are moving in a magnetic field, the Lorentz force will act on them.

It can be established with the aid of the left-hand rule that the free electrons will be displaced to the end  $A$  of the wire. The resulting voltage between the ends of the wire  $A$  and  $B$  will set up an electric force in it,  $F_{\text{el}}$ , to compensate for the Lorentz force  $F_{\text{Lor}}$ . Hence, the flow of electrons to  $A$  will stop as soon as  $F_{\text{el}} = F_{\text{Lor}}$ . Since  $F_{\text{el}} = Eq = Uq/l$  and  $F_{\text{Lor}} = Bvq \sin \alpha$ , we have  $Uq/l = Bvq \sin \alpha$ ; whence

$$U = Bvl \sin \alpha$$

Since the voltage across the terminals of an open circuit is equal to the emf, the formula for the induced emf established in a conductor moving in a magnetic field is

$$\mathcal{E}_{\text{ind}} = Bvl \sin \alpha \quad (26.5)$$

Note that the nonelectric forces responsible for the emf in this case are the magnetic forces acting on the free electrons

in the conductor. In this case, the electrons in the moving wire constitute a "convection current" that produces magnetic effects even though this current is caused by the mechanical motion of the whole wire and not in the usual way by

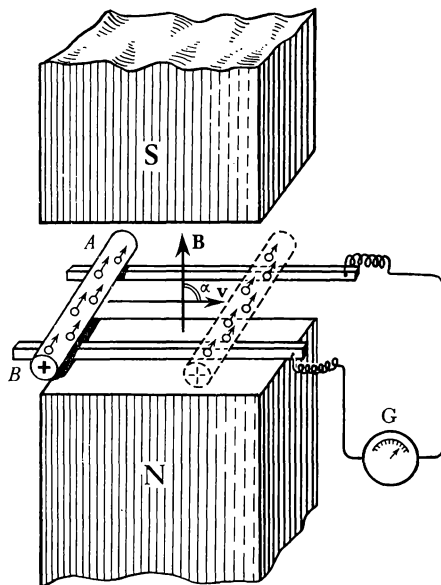


Fig. 26.1 Motion of conductor  $AB$  in magnetic field causes displacement of free electrons to end  $A$ .

the propulsion of the electrons by an external source of emf. The interaction of the magnetic field produced by this convection current of electrons and the surrounding magnet-

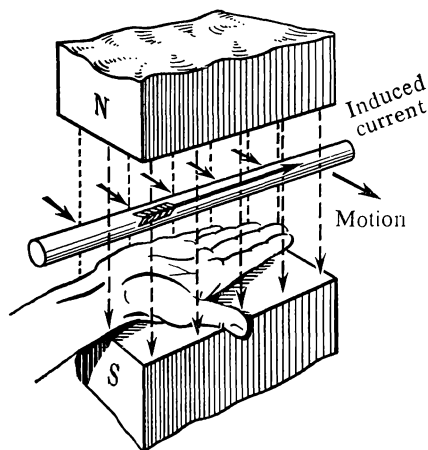
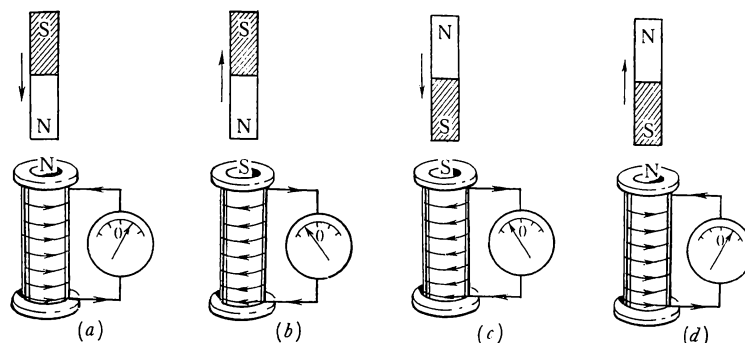


Fig. 26.2 Right-hand rule.

ic field through which it passes is such as to propel the electrons along the wire. If the conductor is connected into a circuit, current will flow in it; this can be inferred from the readings of the galvanometer,  $G$ .

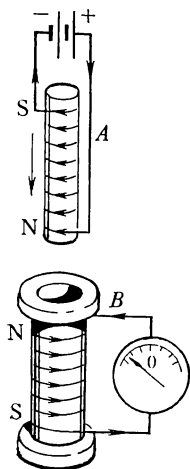
The direction of the induced current appearing in a conductor moving in a magnetic field can be found with the

**Fig. 26.3** When magnet moves with respect to coil, current is induced in it setting up magnetic field with poles shown in coil's centre.



aid of the *right-hand rule* (Fig. 26.2): if one's right hand is placed along the conductor so that the lines of magnetic induction enter the palm and the outstretched thumb points in the direction of the conductor's motion, then the four outstretched fingers will indicate the direction of the induced current in the conductor.

**Fig. 26.4** When coils  $A$  and  $B$  move with respect to one another, current is induced in  $B$ .



## 26-4 Faraday's Induction Experiments

Let us discuss Faraday's experiments that helped him to discover the phenomenon of electromagnetic induction.

*First experiment.* Take a solenoid connected to a galvanometer (Fig. 26.3) and start inserting a permanent magnet into it. The galvanometer's pointer will be seen to deflect as the magnet moves. As soon as the magnet stops, the galvanometer's pointer returns to zero. The same happens when the magnet is withdrawn from the solenoid or when the solenoid is put onto the magnet at rest. Such experiments prove that induced current appears in the solenoid only in the course of a relative displacement of the solenoid and the magnet.

*Second experiment.* We insert into solenoid  $B$  a current-carrying coil  $A$  (Fig. 26.4). In this case, too, the induction current in the solenoid  $B$  is found to appear only in the course of the relative displacement of the solenoid  $B$  and the coil  $A$ .

*Third experiment.* Insert coil  $A$  into solenoid  $B$  and fix them in place (Fig. 26.5). There will be no current in the solenoid. But the moment the current in coil  $A$  is switched on and off, induction current appears in the solenoid. The same happens when the current in coil  $A$  is increased or decreased by means of a rheostat  $R$ .

In future we shall call the circuit of coil  $A$  *primary circuit* and the circuit of solenoid  $B$  in which the induced current appears *secondary circuit*. The same adjectives will be applied to the coils.

*Fourth experiment.* Connect the primary coil into an ac power source and an incandescent lamp across the secondary coil (Fig. 26.6). The lamp burns as long as alternating current flows in the primary coil.

It can be easily seen that the common feature of these experiments is the variation of magnetic field in the solenoid. It is this variation that is responsible for the induced current.

Let us now find out whether any variations of magnetic field in the space around a closed loop induce current in it. Take a plane loop connected to a galvanometer (Fig. 26.7). Place a magnet close to the loop so that its induction lines do not penetrate the loop but run along its plane. When the loop or the magnet are displaced in the plane of the diagram, the galvanometer's pointer stays still. If, however, the loop is rotated about the axis  $OO'$  (Fig. 26.7b), an induced current appears in it.

The experiments described above justify the following conclusions: an induced current (and induced emf) appears in a closed loop only in the case of a variation of the magnetic flux passing through the surface bounded by the loop.

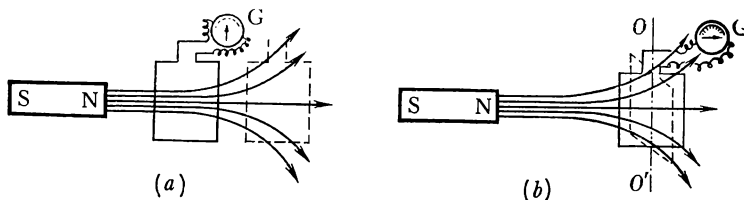


Fig. 26.5 When circuit of coil  $A$  is switched on or off by switch  $S$  or when resistance  $R$  is changed, current is induced in coil  $B$ .

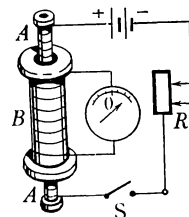


Fig. 26.6 When ac current is fed to primary coil, current is induced in secondary and lamp lights up.

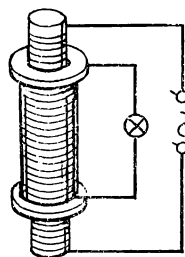


Fig. 26.7 (a) When loop moves, magnetic flux through it remains unchanged since lines of force lie in loop's plane—there is no current; (b) when loop rotates about axis  $OO'$  magnetic flux through it increases—current is induced.

## 26-5 Lenz's Law

Induced current sets up its own magnetic field. Lenz established the relation between the direction of the induced current in a loop and the magnetic field that induced it.

In the experiment depicted in Fig. 26.3 the induced current in the solenoid produces a magnetic field with poles as indicated by letters in the hole of the solenoid. If we consider all four cases of interaction between the magnetic poles of the solenoid and the magnet depicted in the figure with regard to the direction of motion of the magnet, we will conclude that the interaction between the poles is always opposed to the motion of the magnet. Lenz succeeded in generalizing this rule to include all cases of electromagnetic induction. The relation established by him is now known as *Lenz's law*: induced emf sets up an induced current in a closed loop directed in such a way that its magnetic field acts against the cause of the induced emf.

When using Lenz's law to find the direction of induced current one should proceed as follows:

- (1) find the cause of the induced current;
- (2) assuming the induced current to oppose this cause, find the direction of the current's magnetic field;
- (3) find the direction of the induced current from the direction of its magnetic field.

Let us cite an example. The cause of the appearance of induced current in the secondary coil (see Fig. 26.5) the moment the primary coil is disconnected is the disappearance of the field in the primary coil. Accordingly, the induced current in the secondary coil will coincide in direction with that of the current in the primary before it was switched off. (Demonstrate that when the primary circuit is switched on, a current of opposite direction appears in the secondary.)

It follows from Lenz's law that the energy of induced current in a conductor is gained at the expense of the energy spent in opposing the magnetic field of the induced current. For instance, if we disconnect the coil's circuit depicted in Fig. 26.3 and calculate the work required to insert and withdraw the magnet from it a definite number of times, and then repeat the experiment with the circuit closed, we will find the work in the latter case to be substantially greater. The explanation is that in the former case there is no intrinsic field in the coil, while in the latter case such a field is present. The extra work in the second case is spent on overcoming the opposition of this field. It is equal to the energy of the current induced in the coil. It can be easily seen that electromagnetic induction can be used to transform mechanical energy into electric, as well as to transmit electric energy from one circuit to another.

In cases when the induced current is the result of some mechanical motion, electric energy is produced at the

expense of mechanical. Such a transformation takes place in *induction generators*, installed in electric power stations. In cases when the induced current appears in the absence of any mechanical motion, the electric energy is transmitted from one circuit to another. Such energy transport is effected in *transformers* (see Section 29-5).

Electromagnetic induction explains diamagnetism. When a substance is placed in a magnetic field every orbiting electron is acted upon by a Lorentz force, which either increases or reduces (depending on the direction of the electron's rotation), the centripetal force acting on the electron. This changes the orbit and the orbiting frequency of the electron, which is equivalent to the increase or decrease in the circular current corresponding to the orbital motion of the electron, the increase occurring when the magnetic fields of the electron's circular currents are directed against the external field and the decrease when they coincide in direction.

Hence, if in the absence of an external field the circular currents of the electrons in a molecule of a diamagnetic substance compensate each other and the molecule has no intrinsic magnetic moment, then in an external field this equilibrium will be destroyed, with the result that a molecular magnetic moment appears, directed against the field. This result, by the way, follows directly from Lenz's law: the variation of circular molecular currents results in an induced current whose magnetic field should be directed against the external field responsible for it.

The diamagnetic effect appears in all substances, but if the molecules of the substance have intrinsic molecular moments which turn in the direction of the external field and thus augment it, the diamagnetic effect is overridden by the more powerful paramagnetic effect and the substance turns out to be paramagnetic.

A strong diamagnetic effect is observed in conditions of superconductivity. When a superconductor is placed in a magnetic field, currents are induced in it just as in an ordinary conductor. However, in this case they are due not to molecular induced currents but to electron currents. In a superconductor the induced currents meet with no resistance and circulate as long as the external magnetic field exists. They prevent the magnetic field from penetrating the superconductor. Superconductors, as is the case with any diamagnetic substance, tend to be pushed out of magnetic fields.

## 26-6 The Magnitude of Induced EMF

If the reader repeats Faraday's experiments, he or she will observe that the deflection of the galvanometer's pointer increases with the increase in the rate at which the magnet or the current-carrying coil are inserted into the solenoid (see Section 26-4). A similar result is obtained when the magnetic field of the primary coil is increased by increasing the current flowing in it.

More detailed studies of this phenomenon produced the result that emf induced in a circuit is directly proportional to the rate of variation in the magnetic flux linkage of this circuit:

$$\mathcal{E}_{\text{ind}} = -\Delta\psi/\Delta t \quad (26.6)$$

Note that when the circuit is made up of one turn, i.e. is a loop, formula (26.6) reads

$$\mathcal{E}_{\text{ind}} = -\Delta\Phi/\Delta t \quad (26.6a)$$

In these formulae  $\Delta t$  is the time over which the variation of the flux linkage,  $\Delta\psi$ , takes place. If  $\Delta t$  is very small, formulae (26.6) give the instantaneous value of induced emf. If  $\Delta t$  is large, the values obtained from the formulae are average values.

The minus sign in the formulae shows that a decrease in the flux linkage ( $\Delta\psi$  is negative) results in an emf that sets up an induced current which increases the flux linkage, and vice versa. Hence, the minus sign shows that in compliance with Lenz's law induced emf should oppose the agent responsible for it.

It follows from formula (26.6a) that the unit for measuring magnetic flux in the SI system can be termed *volt-second*, since

$$|\Delta\Phi| = |\mathcal{E}_{\text{ind}}\Delta t|, \quad 1 \text{ Wb} = 1 \text{ V} \cdot \text{s}$$

## 26-7 Solenoidal Electric Field and Its Relation to Magnetic Field

The explanation for the appearance of induced emf in a straight conductor moving in a magnetic field (Section 26-3) involved a Lorentz force acting on mobile charge carriers. However, this explanation fails in the case of emf induced in the secondary circuit with the primary remaining at rest with respect to it (the fourth experiment in Section 26-4), because a magnetic field does not act on static charges.

We recall that an electric field can act on static charges. Is it, then, responsible for the current induced in the secondary? If this is indeed the case, where does this electric field come from? The explanation is that a variable magnetic field can set up an electric field which induces current in a closed circuit.

The first to advance this explanation was James Clark Maxwell. Elaborating this idea he created the theory of the electromagnetic field, which was subsequently confirmed by numerous experiments. According to Maxwell's theory a variable magnetic field always sets up in its vicinity in space an electric field with closed lines of strength, no matter whether a substance is present or not.

The straight lines in Fig. 26.8 depict a magnetic field with an induction  $B$  and the rings the generated electric field  $E$ . If there is a conductor there, with the lines of the electric field strength passing in it, a current will flow in it. For example, when the magnet is withdrawn from the coil in Fig. 26.3*d* an electric field depicted in Fig. 26.8*b* appears, setting up a current in the solenoid. (Try to explain the appearance of a current in the cases shown in Fig. 26.3.)

The lines of force of the electric and magnetic fields in Fig. 26.8 are seen to be arranged at right angles. It has been established that at any point in space the magnetic field strength (or induction) vector is perpendicular to the vector electric field strength set up by it. This is why maximum emf in a straight conductor is produced when the conductor moves at right angles to the induction lines of the magnetic field.

## 26-8 Eddy Currents

Take a coil with a protruding core made of a soft ferromagnetic substance and place a metal object on top of it. If the coil is connected to an ac power source, the object will soon become hot.

We replace the object with an aluminium ring around the core and again connect the coil to the power source (Fig. 26.9). If one holds the ring, it will become hot; if it is released, it slides off the core. The explanation for these phenomena is that the variable magnetic field around the core sets up an electric field which induces large currents in the object and in the ring, because their resistances are very small. The ring slides off the core because the current induced in the ring is opposite in direction to the current in the coil and such currents repel each other.

Fig. 26.8 Electric field due to variation of magnetic field: magnetic induction (a) increases and (b) decreases.

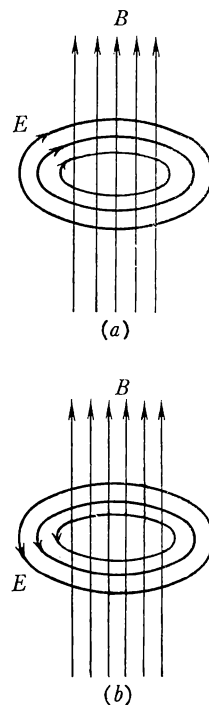
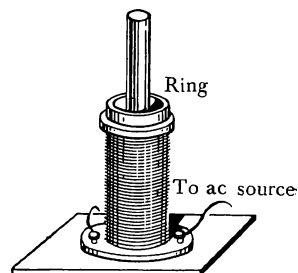
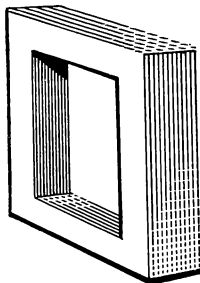


Fig. 26.9 Aluminium ring placed on core jumps off when coil is connected into ac circuit; if held, it is heated by eddy currents.

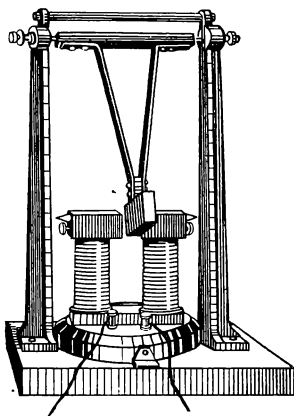




**Fig. 26.10** To reduce eddy currents, core of transformer is made of separate insulated sheets of transformer steel.



**Fig. 26.11** Eddy currents hamper motion of pendulum between poles of electromagnet.



Currents induced in solid metal bodies placed in variable magnetic fields and closed inside such bodies are termed *eddy*, or *Foucault, currents* (in honour of the French physicist Jean Bernard Foucault (1819-1868) who studied these currents).

The rotor of an electric motor and the core of a transformer operate in variable magnetic fields and therefore eddy currents should circulate in them. The energy spent on inducing eddy currents is transformed into the internal energy of the rotor and the core, that is, into heat. Since such parts are made of ferromagnetic materials, losses due to eddy currents are not the only losses, others being due to hysteresis. To reduce the harmful effects of eddy currents parts designed to work in variable magnetic fields are made of separate plates insulated from each other (Fig. 26.10).

Note that the specific resistance of ferrites is very great. Therefore virtually no eddy currents are induced in them and this appreciably reduces the losses. Since the losses due to hysteresis in ferrites are also quite small, their use results in a noticeable improvement in the efficiency of various devices, for instance, transformers.

If eddy currents are the result of the motion of a body in a magnetic field, then, in compliance with Lenz's law, these currents should arrest the motion of the body. The arresting action of the eddy currents can be demonstrated with the aid of the following experiment.

If a plate (Fig. 26.11) is made to swing first in the absence of a current and then with a current flowing in an electromagnet, the oscillations will be seen to stop almost instantly when the current starts to flow. The plate appears to stick as if in a thick fluid. The arresting action of eddy currents is used in instruments to damp the vibrations of the measuring mechanism.

Modern technology makes use of eddy currents for heat treatment of parts and for melting alloys in induction furnaces.

## 26-9 Self-Induction and Self-Induced EMF

We recall that the intrinsic magnetic field in a dc circuit varies at the moments the circuit is switched on and off, as well as during the time the current in it changes. This means that at such moments an induced emf should appear in the circuit. The appearance of an induced emf in a circuit caused by the variation of the magnetic field of the

current flowing in the circuit is called *self-induction* and the emf is called *self-induced emf*.

Let us examine what happens when the circuit is closed. We have an open circuit (Fig. 26.12) consisting of a power supply B, a switch S, a lamp M and a coil C with a ferromagnetic core, all connected in series. When the circuit is closed, the lamp, after some delay, lights up. The explanation is that a substantial emf is self-induced in the circuit, which, in compliance with Lenz's law, obstructs the rapid rise of the current in the circuit. The growth of current in such a circuit is depicted in Fig. 26.13, where  $I_0$  is the dc current.

Note that the energy of the power source spent on overcoming the self-induced emf is accumulated in the magnetic field of the circuit, mainly inside the coil with core C. (Why?) When the current in the circuit attains its constant value, the energy of the magnetic field of the circuit no longer changes. The magnitude of the energy of the magnetic field of the circuit depends not only on the current but also on the type of the circuit, that is, on its self-inductance  $L$ . Magnetic energy is especially great in powerful electromagnets.

To observe self-induction at the moment the circuit is switched off a circuit like that shown in Fig. 26.14 can be assembled. When this circuit is switched off by means of switch S, the circuit consisting of the coil C and the lamp M remains closed. When the current in coil C starts falling off rapidly (Fig. 26.15), a self-induced emf appears in it which retards the decline in the current. In this situation the coil acts for a short time as a power source maintaining current in the lamp M. At the moment the circuit is broken the current in the lamp falls off to zero and with a jump changes direction, rising to a value which may far exceed the current in the lamp before the circuit was disconnected. This may cause the lamp to flash brightly at the moment the circuit is broken or even to burn out.

Self-induction is responsible for sparks appearing at the place where contact is broken. If the circuit contains powerful electromagnets storing large amounts of magnetic energy, the spark may develop into an arc discharge and ruin the switch. To break such circuits at power stations oil switches and other precautionary measures are used.

Deduce a formula for calculating self-induced emf. Since formula (26.6) holds for all types of emf's:

$$\mathcal{E}_{\text{ind}} = - \frac{\Delta \psi}{\Delta t}$$

Fig. 26.12 Self-induced emf retards increase in current flowing through lamp when circuit is closed.

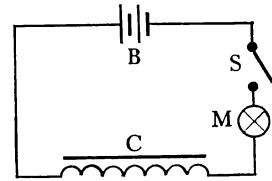


Fig. 26.13 Current rising in circuit of Fig. 26.12 after circuit has been closed.

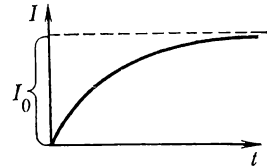


Fig. 26.14 When circuit is disconnected by switch S, lamp at first flashes brightly and then goes out.

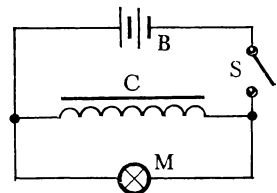
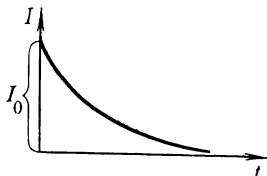


Fig. 26.15 Current falling in coil C (Fig. 26.14) after circuit had been broken by switch S.



and since  $\psi = LI$ , it follows that

$$\mathcal{E}_{\text{self}} = - \frac{\Delta(LI)}{\Delta t}$$

whence

$$\mathcal{E}_{\text{self}} = -L \frac{\Delta I}{\Delta t} \quad (26.7)$$

The self-induced emf in a circuit is directly proportional to the current variation rate in the circuit.

## 26-10 The Energy of a Magnetic Field

It was stated in the preceding section that the energy of the magnetic field of a circuit depends on the current flowing in it and on its shape. Let us dwell on this dependence. We recall that the energy of the magnetic field of a circuit,  $E_{\text{mag}}$ , is equal to the work spent on overcoming the self-induced emf, which appears at the moment the circuit is closed. If the average emf in this case was  $\mathcal{E}_{\text{self}}$  and if the charge that passed through the circuit during the time  $\Delta t$  the current rose was  $q$ , then the work spent on overcoming the self-induced emf would be  $\mathcal{E}_{\text{self}} q$ . In that case

$$E_{\text{mag}} = -\mathcal{E}_{\text{self}} q$$

The minus sign means that the charges move against the self-induced emf. Since  $\mathcal{E}_{\text{self}} = -L\Delta I/\Delta t$ , it follows that

$$E_{\text{mag}} = L \frac{\Delta I}{\Delta t} q = L\Delta I \frac{q}{\Delta t}$$

Since the current in the circuit rises from 0 to  $I$ , it follows that  $\Delta I = I - 0 = I$  and  $q/\Delta t$  is the average current during  $\Delta t$ . Assuming the average current to be equal to  $I/2$  and substituting the values of  $\Delta I$  and  $q/\Delta t$  into the above relation, we find the formula for the magnetic energy of the circuit carrying a current  $I$ :

$$E_{\text{mag}} = LI \frac{I}{2}, \quad \text{or} \quad E_{\text{mag}} = \frac{LI^2}{2} \quad (26.8)$$

The energy of the magnetic field of a circuit is directly proportional to its inductance  $L$  and to the square of the current flowing in it. Since the inductance of a solenoid containing a ferromagnetic core is especially large, great magnetic energy is stored in a circuit containing electromagnets. Since  $LI = \psi$ , we obtain

$$E_{\text{mag}} = \frac{\psi I}{2}, \quad \text{or} \quad E_{\text{mag}} = \frac{\psi^2}{2L} \quad (26.8a)$$

part three

# **Oscillations and Waves**

# Mechanical Oscillations and Waves

## 27-1 Oscillatory Motion

There are numerous periodic processes in nature and technology based on oscillations of various types and on the waves produced by them. Such processes include sound phenomena, the operation of clock mechanisms, alternating current flowing in circuits, electromagnetic oscillations, etc.

The nature of oscillations is diverse and yet measures have been devised to describe them which have turned out to be equally applicable to all types of oscillations regardless of their nature.

The easiest way to elucidate the physical meaning of these quantities and to make oneself familiar with the mathematical apparatus developed to describe oscillatory processes is to use the example of mechanical oscillations. These have the advantage that we are able to visualise them. Therefore it pays to begin the study of oscillatory and wave processes by looking at the special features of mechanical oscillations.

Figure 27.1 depicts bodies involved in various oscillations: (1) a swinging pendulum, (2) an oscillating liquid, (3) an oscillating spring and load, (4) a vibrating string. All bodies in Fig. 27.1a are in a state of stable equilibrium.

For them to leave the state of stable equilibrium an excess energy has to be transmitted to them at the expense of work done by some external force. Then they will assume the position shown in Fig. 27.1*b*. If the bodies are now left to move unobstructed, they will start oscillatory motion and pass positions (c) (d) (e) in succession, returning to (b) to repeat the cycle in the same order.

If we observe the motion of some point of a body, we shall see that it moves along the same trajectory in opposite directions. Since the nature of motion of all points of a body is similar, it is convenient to study oscillatory processes using the oscillations of one point of a body as an example. We recall once more that the most important idea about oscillatory motion is that it repeats itself after equal time intervals  $T$ , or that it is *periodic*.

*Mechanical oscillation* is the term for a periodic motion of a particle along a path which the particle covers in turn in opposite directions. Note that at each point of the path the velocities of the particle moving in opposite directions are equal in magnitude.

*Complete oscillation* is the term for one completed cycle of oscillatory motion, after which it is repeated in the same order.

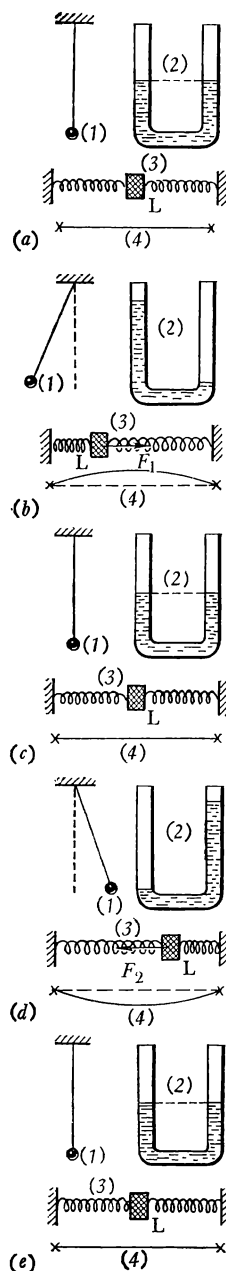
## 27-2 Conditions for Appearance of Oscillations

Let us find the conditions for oscillations to start and continue for some time.

The *first condition* is that the particle should possess energy (kinetic or potential) in excess of its energy in the stable equilibrium position (see Section 27-1).

We see the second condition being established if we observe the motion of the load  $L$  in Fig. 27.1. In position (b) a force  $F_1$  acts on  $L$  in the direction of its equilibrium position (see Fig. 27.1*a*). Acted upon by this force the load moves towards the equilibrium position at a gradually increasing speed  $v$ , the force  $F_1$  decreasing in the meantime and vanishing altogether the moment the load reaches this position (Fig. 27.1*c*). At this moment the speed of the load is at its maximum and, after passing the equilibrium position, it continues to move to the right. This is accompanied by the appearance of an elastic force  $F_2$  which arrests the motion of  $L$ , finally stopping it (position (d) in Fig. 27.1). In this position force  $F_2$  is at its maximum, and acted upon by this force load  $L$  starts moving to the left. In the equilibrium position (Fig. 27.1*e*)  $F_2$  vanishes and  $v_2$  attains its maximum

Fig. 27.1 Examples of oscillations of various bodies.



value, and so the load continues on its way to the left until it reaches position (b) in Fig. 27.1. The process is then repeated in the same order.

Hence the oscillations of a load are due to the existence of a force and of the inertia of the load. A force applied to a particle always directed towards the stable equilibrium position is termed a *restoring force*. In the stable equilibrium position the restoring force is zero, increasing with the particle's displacement from this position.

Hence, the *second condition* for the appearance and the continuation of oscillations of a particle is the action of a restoring force on the particle. Recall that such a force always appears when a body is displaced from a position of stable equilibrium.

Ideally, in the absence of friction and of resistance of the medium the total mechanical energy of an oscillating particle remains constant, since such oscillations involve only transformations of kinetic energy into potential, and vice versa. Such oscillations can continue for an indefinite time.

If the oscillations of a particle take place in the presence of friction and the resistance of the medium, then the total mechanical energy of the particle gradually diminishes, the amplitude decreases and after some time the particle stops in the stable equilibrium position.

In some cases the energy losses of the particle are so great that, being displaced by an external force from the stable equilibrium position, it loses its excess energy before it reaches the equilibrium position. In this case there will be no oscillations. Hence, the *third condition* for the appearance and the continuation of oscillations: excess energy gained by the particle in the course of its displacement from the stable equilibrium position should not be completely spent on its return to that position.

It follows that the number of cycles completed by a particle once displaced from its equilibrium position will be the greater the less energy it spends on overcoming resistance in each cycle.

### 27-3 Classification of Oscillatory Motion Based on the Forces Acting on the Source

It was established in the preceding section that there is always a restoring force acting on an oscillating particle. The term used for oscillations of a particle (a body) resulting only from the action of a restoring force is *natural os-*

*cillations* of the particle (or the body). Note that natural oscillations do not exist in the real world since there is always a force of resistance of the medium (friction force) acting on an oscillating body. The effect of the force of resistance is to retard somewhat the whole oscillatory process, that is, to increase the time of one complete cycle and to gradually decrease the amplitude of oscillation. The oscillations of a particle acted upon by the force of resistance of the medium and by a restoring force are called *free oscillations*.

It can be easily seen that the difference between free and natural oscillations is the smaller the less the resistance of the medium in which such oscillations take place. Hence, natural oscillations may be regarded as the limiting case of free oscillations, when the resistance of the medium tends to zero.

Consider another type of oscillations. When an engine running rhythmically, for instance an electric motor, stands on the floor, the floor vibrates. The hull of a motor boat and the fuselage of a plane also vibrate when their engines are running. These vibrations are the result of periodic external action.

The oscillations of a body resulting from a periodic force acting on it are called *forced oscillations*. In this case the forces acting on the oscillating body include, in addition to a periodic external force, a resistance force and a restoring force as well. Accordingly, the nature of forced oscillations depends on the relation between those forces.

#### 27-4 The Parameters of Oscillatory Motion

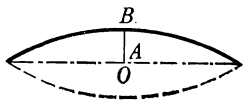
There are some characteristic quantitative features of an oscillatory motion which make it possible to distinguish it from other oscillations and which remain constant provided certain conditions are fulfilled. They are termed *parameters of oscillatory motion*.

The first such feature is its periodicity. The term for the quantity  $T$  characterizing the periodicity of oscillatory motion is *period of oscillation*. The measure for the period of oscillation is the time of one complete oscillation, usually expressed in seconds.

A related feature of oscillatory motion is its frequency (repetition rate). The term used for the quantity  $f$  characterizing the repetition rate of the oscillatory motion is *frequency of oscillation*. The measure for the frequency of oscillations



Fig. 27.2 Amplitude  $A$  of oscillations of point  $O$  is  $OB$ .



of a body is the number of oscillations completed per unit time:

$$f = 1/T \quad (27.1)$$

(Ask yourself why the frequency of oscillation is expressed by formula (27.1).)

We now deduce a unit for measuring frequency:

$$f = 1/T, \quad f = 1/1 \text{ s} = 1 \text{ s}^{-1} = 1 \text{ Hz (hertz)}$$

The accepted unit for measuring frequency is the *hertz*. One hertz is the frequency of oscillation of a body making one complete oscillation in one second.

The second characteristic feature of an oscillatory motion is its *amplitude*. Amplitude is the term for the maximum deflection  $A$  of an oscillating particle from its stable equilibrium position (Fig. 27.2).

Hence, there are two parameters of oscillatory motion: the period  $T$  (or the frequency  $f$  related to the period by means of formula (27.1)) and the amplitude  $A$ .

Oscillations of a particle whose amplitude does not change with time are termed *undamped oscillations*, and those with a gradually diminishing amplitude are termed *damped*. Recall that free oscillations are damped. One should note that undamped oscillations are not necessarily natural oscillations. Forced oscillations may also take place with a constant amplitude. Since the resistance of the medium transforms mechanical energy into internal energy, the energy of the oscillating particle should be periodically replenished at the expense of the work of some periodic force. An example of such oscillations are the oscillations of the balance in a watch whose energy is replenished at the expense of the energy of a wound up spring. Note that the energy of oscillatory motion is determined by its parameters and by the mass of the oscillating particle. The theory of oscillations proves that the excess energy of an oscillating particle is directly proportional to its mass, to the square of its amplitude and to the square of its frequency.

There is one additional point to be noted. The frequency of forced oscillations coincides with the frequency of the driving force (it may also be a multiple of the force's frequency).

### 27-5 Quantities Characteristic of the Instantaneous State of an Oscillating Particle

The parameters  $T$  (or  $f$ ) and  $A$  remain constant in the case of undamped oscillations. However, there are quantities that vary continuously in the course of the oscillations of a particle. Their numerical values depend on time. Therefore they may be said to characterize the instantaneous state of the oscillating particle.

The first such quantity is the one characterizing the position of the oscillating particle with respect to its stable equilibrium position at a chosen instant, the *displacement* (we denote it  $x$ ). The measure of displacement is the distance of the oscillating particle from its stable equilibrium position at the chosen instant. To make the numerical value of the displacement unique a sign is attributed to it. For instance, if the displacement of a load  $L$  to the right from its stable equilibrium position in Fig. 27.1 is assumed to be positive, then its displacement to the left will be negative.

It can be easily seen that the amplitude  $A$  is numerically equal to the maximum displacement  $x_{max}$  of an oscillating particle from its stable equilibrium position:

$$A = |x_{max}| \quad (27.2)$$

The second important parameter, characterizing both the position and the direction of motion of an oscillating particle at a specific instant, is the *phase of oscillation* (we denote it  $\varphi$ ). The measure of the phase is an abstract number equal to the ratio of the time that elapsed from the start of the oscillations through to the specified instant to the period.

Although the value of the phase can be quite great, it is usually expressed as a proper fraction, the whole periods elapsed from the start of oscillations being discarded since after a complete period the process is repeated in the same order.

To make the value of the phase unique it should be decided which position of the oscillating particle will be taken as the initial position for registering the phase. We can clarify this by using a pendulum as an example (Fig. 27.3).

Assume the origin for measuring the phase to be the position of equilibrium of the pendulum swinging from right to left. In this case the variations of phase during the oscillations of the pendulum will be expressed by the numbers shown in Fig. 27.3. If we are told that the phase of such a

Fig. 27.3 Variation in phase in oscillations of pendulum.

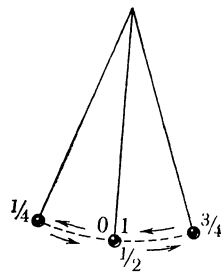
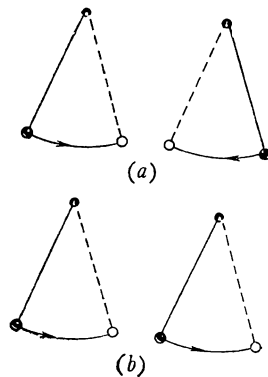


Fig. 27.4 Oscillations of pendulums: (a) in-phase; (b) out of phase.



pendulum is  $1/2$ , this means that it is in the equilibrium position and moving to the right.

Note that the difference between displacement and phase is not only that the latter determines the direction of motion. The displacement of an oscillating particle depends on its amplitude of oscillation, but the phase does not. (Can two identical pendulums have identical phases but different amplitudes?)

The phase helps to differentiate between the oscillations of particles whose periods and amplitudes are identical. Figure 27.4*a* shows two identical pendulums which simultaneously start oscillating from positions shown in the figure. Their periods and amplitudes are identical, but they swing in opposite directions. If we measure the phases of both oscillations from the same initial moment, we are able to express the difference between the oscillations in terms of *phase difference*. Suppose that both pendulums are held in the extreme left position. If we release the right pendulum and subsequently, when it reaches the extreme right position, release the left pendulum, they will continue to swing as shown in Fig. 27.4*a*. Since the left pendulum started its oscillations from the same initial position as the right, but half a period later, the latter is said to be leading the former in phase by  $1/2$ . (Consider why the right pendulum may be said to be lagging behind the left pendulum in phase by  $1/2$ .) Hence, the pendulums in Fig. 27.4*a* oscillate with a phase difference (also termed *phase shift*) of  $1/2$ , and in Fig. 27.4*b* with a zero phase difference. The following statement is essential for the future discussion: the phase difference between two oscillations of identical period (or frequency) remains constant during the time the oscillations continue. The text below refers to such oscillations.

The oscillations of two particles with a zero phase difference are said to be *in-phase*. Specifically, if both particles continuously move in the same direction their phases are equal. Two particles oscillating with a phase difference of  $1/2$  are said to be in *opposite phases*. Specifically, if oscillating particles always move in opposite directions, their phases are opposite.

Other characteristics of the instantaneous state of an oscillating particle besides the displacement and the phase are its velocity  $v$  and acceleration  $a$ , since these quantities also continuously vary with time. The velocity is at its maximum in the stable equilibrium position and becomes zero in the extreme positions. The acceleration is zero in the equilibrium position and is at its maximum in the extreme positions (explain why).

## 27-6 Harmonic Oscillations

In cases when the role of the restoring force is played by the resultant of an elastic force and the force of gravity, the parameters of oscillatory motion can be related to the parameters of circular motion of a particle.

To find this relation we do the following experiment. Place a turntable in front of a screen and place on it a rod with a ball  $B$  at its end (Fig. 27.5). Arrange the light so

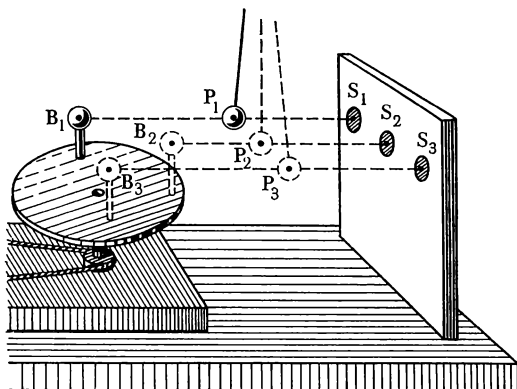


Fig. 27.5 Motion of shadow of particle moving uniformly in a circle is oscillatory, similar to that of pendulum.

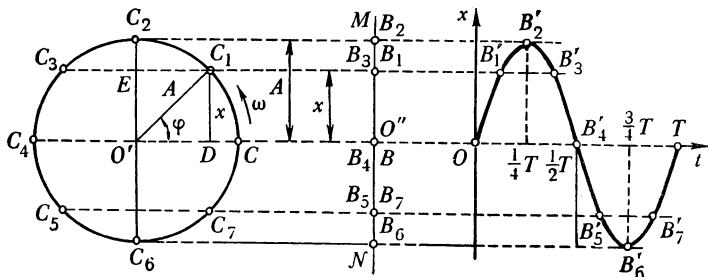
that there is a sharp shadow  $S$  from the ball on the screen. Place a pendulum  $P$  between the screen and the ball so that its shadow coincides with that of the ball on the screen. The pendulum can be made to swing in such a way that one can rotate the disk at a constant speed and the shadows of the ball  $B$  and of the pendulum  $P$  on the screen continuously coincide. This proves that the projection of the ball on the screen performs exactly the same oscillatory motion as pendulum  $P$ .

Hence, the projection of a particle that is moving in a circle of radius  $A$  at constant circular speed with a period  $T$  on one of its diameters is equivalent to the oscillations of a particle with amplitude  $A$  and period  $T$ . This makes it possible to study the special features of oscillations by making use of the motion of the projection of such a particle along the diameter of the circumference.

Let point  $C$  in Fig. 27.6 move in a circle of radius  $O'C = A$  with a constant angular velocity  $\omega$  and make a complete cycle in the time  $T$ . In such a case the projection of point  $C$  on the straight line  $MN$  will oscillate with an amplitude  $A$  and a period  $T$ . Start marking time from the instant the mobile radius is in position  $O'C$  and the oscillating point

is in position  $O''$ . Let the radius turn by an angle  $\phi = \omega t$  in time  $t$  with the projection of its end moving along the straight line  $MN$  for a distance  $x = DC_1 = O''B_1$ . The positions of the end of the mobile radius corresponding to successive equal time intervals are marked by points on the circle, similar points on the straight line  $MN$  making the simultaneous positions of the oscillating point.

Fig. 27.6 Relation between motion of particle in circle and motion of its projection along the circle's diameter; graph of harmonic oscillations is to right.



The displacement of the oscillating point  $B$  from the equilibrium position may be found from triangle  $O'C_1D$ :

$$x = A \sin \phi \quad (27.3)$$

or

$$x = A \sin \omega t \quad (27.3a)$$

In these formulae  $\phi$  is the *phase angle* or simply *phase*. It is expressed in radians. The quantity  $\omega$  as applied to oscillatory motion is termed *angular*, or *cyclic*, *frequency*. Since the formulae expressing angular velocity in the case of uniform circular motion of a point are

$$\omega = 2\pi/T, \text{ or } \omega = 2\pi f \quad (27.4)$$

we obtain for the phase angle

$$\phi = (2\pi/T) t, \text{ or } \phi = 2\pi f t \quad (27.5)$$

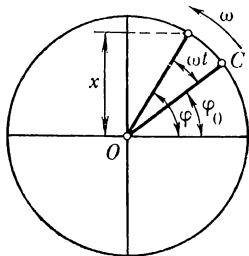
It follows from these formulae that the only difference between the phase expressed in radians and in fractions of a period  $t/T$  is the presence of a constant factor  $2\pi$ .

The time can be marked from any moment, for instance from the moment the point occupies position  $C$  in Fig. 27.7. In this case the initial position of the point is determined by the angle  $\phi_0$ , known as the *initial phase*. In this case the phase of the oscillation can be expressed in the form

$$\phi = \phi_0 + \omega t \quad (27.6)$$

$$\phi = \phi_0 + (2\pi/T) t, \text{ or } \phi = \phi_0 + 2\pi f t \quad (27.6a)$$

Fig. 27.7 If initial phase is  $\phi_0$ , then  $\phi = \phi_0 + \omega t$ .



Oscillations described by formula (27.3) are often termed *sinusoidal*. In physics the term used for oscillations obeying the sine law is *harmonic*. Specifically, oscillations resulting from the action of only one restoring force proportional to displacement are harmonic, that is, if the expression for the restoring force is

$$F_r = -kx \quad (27.7)$$

and there are no other forces acting on the particle, its oscillations will be harmonic.

### 27-7 The Equation for Harmonic Oscillations and Its Graph

The term used for the formula expressing the time dependence of the displacement of the oscillating particle is the *equation for harmonic oscillations*. Hence, formula (27.3) can be said to be an equation for harmonic oscillations. A more general equation for harmonic oscillations can be obtained if the values of  $\varphi$  from formulae (27.6) and (27.6a) are substituted for  $\varphi$  in (27.3):

$$x = A \sin(\varphi_0 + \omega t) \quad (27.8)$$

$$x = A \sin\left(\varphi_0 + \frac{2\pi}{T} t\right), \quad \text{or} \quad x = A \sin(\varphi_0 + 2\pi f t) \quad (27.8a)$$

Oscillations with the initial phase equal to  $\pi/2$  obey the cosine law:  $\sin(\pi/2 + \omega t) = \cos \omega t$ . Of course, these oscillations are also harmonic.

The graph for harmonic oscillations is a sinusoid and it is plotted as follows. Take point  $O$  on the extension of the straight line  $O'O''$  (see Fig. 27.6) and make it the origin. Mark time  $t$  on the horizontal axis and displacement  $x$  on the vertical axis. Marking on the horizontal axis points  $1/8T$ ,  $1/4T$ , etc. also mark the corresponding displacements on the vertical axis:  $B'_1$ ,  $B'_2$ , etc.

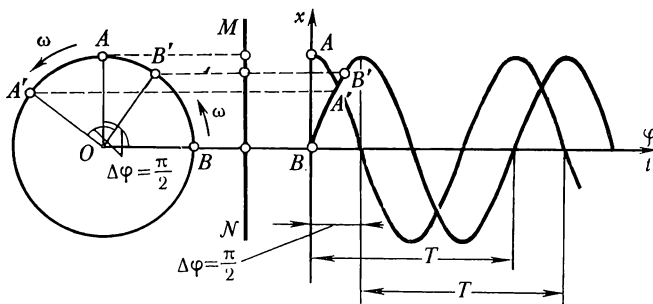
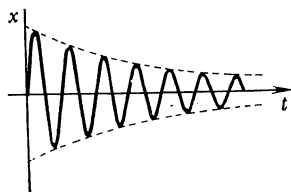
Joining the points  $B'_i$  by a smooth curve, we obtain the graph of the harmonic oscillations of the particle. Figure 27.6 shows the graph for one period  $T$ . Each new period will add an identical section to the graph.

Figure 27.8 shows graphs of two harmonic oscillations with identical periods and amplitudes, but with a phase difference of  $\pi/2$ . The oscillating particle with the left-hand graph (sine curve for  $A$ ) leads in phase the second particle with the right-hand plot (sine curve for  $B$ ) by  $\pi/2$ . It can be seen from Fig. 27.8 that the mobile radius  $OA'$  leads the mobile radius  $OB'$  by  $\pi/2$ .

In the case of damped oscillations the period remains constant while the amplitude decreases continuously. The graph of damped oscillations is shown in Fig. 27.9. Hence,

Fig. 27.8 Oscillation whose plot is to left leads other oscillation by phase  $\Delta\varphi = \pi/2$  (initial phase of oscillation of  $A$  is  $\pi/2$  and of  $B$  is zero).

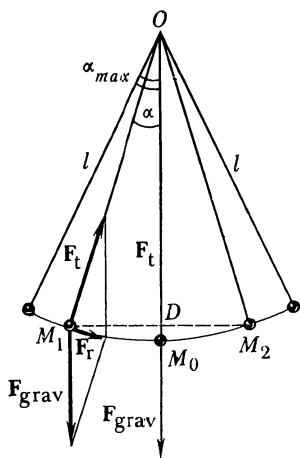
Fig. 27.9 Graph of damped oscillations.



in the case of free oscillations the oscillatory process is not strictly repetitive and can be regarded as harmonic only as an approximation.

## 27-8 The Simple Pendulum

Fig. 27.10 Simple pendulum.



If we observe the oscillations of pendulums of various length, we quickly establish that their periods are related to their lengths. However, in the case of real pendulums it is not always obvious what the actual length is. For instance, by moving the disk on the pendulum of a wall clock one can change the period of its oscillations although the length of the rod remains unchanged. To bypass this difficulty we consider first the case of the so-called simple pendulum. The definition of the length of such a pendulum presents no problems.

*Simple pendulum* is the term used for a particle suspended on a weightless and inextensible thread. A small heavy ball, for example of lead, suspended on a thin long inextensible thread is a good model of a simple pendulum. First, let us deal with the problem of whether the oscillations of a simple pendulum can be regarded as harmonic. To this end we must find the dependence of the restoring force  $F_r$  on the displacement.

Let a simple pendulum (Fig. 27.10) of length  $l$  be deflected from its equilibrium position  $OA$  by an angle  $\alpha$  and be in the position  $OB$ . The restoring force  $F_r$  is equal to the vector sum of the force of gravity  $F_{grav}$  and the tensile strength of the thread  $F_t$  and is directed along the tangent to the arc  $AB$ . The condition of similarity of triangles

$BNM$  and  $BOD$  yields

$$\frac{F_r}{F_{\text{grav}}} = \frac{BD}{BO}$$

whence

$$F_r = F_{\text{grav}} \frac{BD}{BO} = F_{\text{grav}} \frac{BD}{l}$$

The pendulum returns to the position of equilibrium along the arc  $BA$ . Thus the displacement  $x$  is equal to the length of the arc  $BA$ . For small angles  $\alpha$  the length of arc  $BE$  is approximately equal to that of chord  $BE$ , and arc  $BA$  is equal to half the chord  $BE$ , that is,  $\widehat{BA} \approx BD$  or  $x \approx BD$ . Therefore for small  $\alpha$ 's we can assume that

$$F_r = -F_{\text{grav}} \frac{x}{l} = -\frac{mg}{l} x \quad (27.9)$$

The minus sign is there because  $x$  and  $F_r$  are always opposite to each other.

The quantities  $m$ ,  $g$  and  $l$  are constants when the pendulum oscillates at a fixed point on the Earth's surface. It follows from (27.9) that the restoring force  $F_r$  is directly proportional to the displacement, that is, is expressed by formula (27.7). One should be reminded that this is true only for sufficiently small angles  $\alpha$ . Note that if  $\alpha_{\text{max}}$  is the maximum deflection angle of the pendulum,  $2\alpha_{\text{max}}$  is the *swing angle*.

To summarise, when the swing angle is small (not exceeding several degrees) the oscillations of a simple pendulum can be regarded as harmonic.

## 27-9 Laws Governing the Oscillations of a Simple Pendulum

We will now examine the factors determining the period of oscillations of a simple pendulum. It can easily be established from an experiment performed with a model of a simple pendulum that its oscillations are damped. We have already seen that the period of a pendulum does not change as the oscillations are attenuated, that is, it is independent of the amplitude (for small swing angles). This property of the pendulum was discovered by Galileo and is known as *isochronism* (uniformity in time). Experiment also shows that the period of oscillations of a pendulum is independent of its mass. (Prove it yourself, using formula (27.9).)

Demonstrate, using formula (27.9), that the period of oscillations of a pendulum depends on its length  $l$ . Since



an increase in  $l$  involves a decrease in the restoring force  $F_r$ , the same is true of the acceleration of the pendulum, and its period increases accordingly. It follows from the same formula that an increase in  $g$  involves an increase in  $F_r$  and therefore a decrease in the period.

The properties of the simple pendulum are formulated in the form of two laws:

(1) for small swing angles the period of oscillations of a simple pendulum is independent both of its amplitude and of its mass;

(2) the period of oscillations of a simple pendulum is directly proportional to the square root of the pendulum's length and inversely proportional to the square root of acceleration of free fall,  $g$ ,

$$T = 2\pi \sqrt{l/g} \quad (27.10)$$

(formula (27.10) is obtained from the theory of oscillations).

Note that the term used for half a complete oscillation is *single oscillation*; for example, the motion of a pendulum from one extreme position to another. Since the period of a single oscillation is  $T_s = T/2$ , the formula for the period of a single oscillation of a simple pendulum is

$$T_s = \pi \sqrt{l/g} \quad (27.11)$$

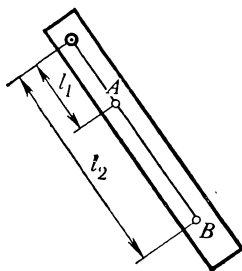
## 27-10 The Compound Pendulum

The laws of oscillations of a simple pendulum can be applied only to oscillations of bodies whose dimensions are small as compared with the distance from the pivot to the centre of mass. The term for all pendulums for which this condition is not satisfied is *compound pendulums*. An example of such a pendulum is given in Fig. 27.11.

The oscillations of a compound pendulum can be thought of as concerted oscillations of numerous interconnected particles, that is, of a multitude of simple pendulums of various length (two of them are shown in Fig. 27.11). This means that formula (27.10) is inapplicable to a compound pendulum. Indeed, the period of oscillations of the ruler depicted in Fig. 27.11 will obviously exceed that of the simple pendulum of length  $l_1$ , but be less than that of the simple pendulum of length  $l_2$ .

In order to apply formula (27.10) the following procedure is adopted. The physical pendulum is made to swing; then the number of oscillations it makes in a specified time is counted. Its period  $T$  is calculated, and finally the length

Fig. 27.11 Compound pendulum.



of a simple pendulum with the same value of  $T$  is computed from formula (27.10).

The term for the length of simple pendulum  $l$  with a period equal to that of the compound pendulum is *reduced length* of the compound pendulum. Hence, formula (27.10), in the case of a physical pendulum, should be rewritten in the form

$$T = 2\pi \sqrt{l_{\text{red}}/g} \quad (27.12)$$

Note that a pendulum with the period of a single oscillation equal to one second is said to be a *second pendulum*. The reduced length of such a pendulum for the latitude of Moscow is 0.99 m.

### 27-11 Practical Uses of Pendulums

The most familiar use of a pendulum is in a clock for timekeeping. This was first accomplished by the Dutch physicist, mathematician and astronomer Christian Huygens (1629-1695).

Since the period of oscillations of a pendulum depends on free fall acceleration  $g$ , a clock which is accurate, say, in Moscow will gain time in Leningrad. To make it run accurately in Leningrad the reduced length of its pendulum should be increased. (Explain this using formula (27.12).)

In geology the pendulum is used for measuring  $g$  at various points on the surface of the Earth. To this end the period at the place of measurement is found from a great number of oscillations and  $g$  is then computed from formula (27.12):

$$g = \frac{4\pi^2 l_{\text{red}}}{T^2}$$

A noticeable deviation of the local value of  $g$  from the normal value for that location is termed *gravity anomaly*. The discovery of anomalies helps in the search for minerals.

It has been established by experiment that a swinging pendulum tends to retain its plane of oscillations. This means that if a pendulum mounted on a turntable is set in motion and then the disk of the turntable is rotated, the plane of oscillations of the pendulum will remain constant with respect to the room. This makes it possible to observe the Earth's rotation about its axis.

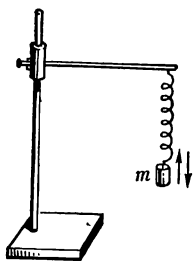
In 1850 J. Foucault set up, in the Pantheon in Paris, a simple pendulum, so that when swinging, its point made a trace in the sand covering the floor. It turned out that

with each new swing the pendulum left a new trace in the sand. Thus Foucault's experiment demonstrated the Earth's rotation about its axis. (Consider what would happen in the absence of the Earth's rotation.)

### 27-12 Elastic Oscillations. Energy Transformation in Oscillatory Motion

The term used for oscillations in which the restoring force is the result of elastic forces is *elastic oscillations*. The oscillations of a spring with the load  $L$  (Fig. 27.1) or the vertical oscillations of the weight  $m$  suspended on a spring (Fig. 27.12) may serve as examples of elastic oscillations.

Fig. 27.12 Elastic oscillations of weight  $m$  suspended on spring.



Since elastic force is proportional to absolute deformation (see Section 13-7), the elastic force of the spring is expressed by formula (27.7):  $F_r = -kx$ , where  $x$  is the displacement of the weight attached to the spring and  $k$  is the force constant of the spring. Hence, natural elastic oscillations of a system are harmonic oscillations. Theory shows that the period of elastic oscillations is determined by the formula

$$T = 2\pi \sqrt{m/k} \quad (27.13)$$

(Demonstrate that (27.13) may be used to obtain the formula for the period of oscillations of a simple pendulum (27.10):  $T = 2\pi \sqrt{l/g}$ , taking into account that the restoring force acting on the pendulum  $F_r = -mgx/l$  is also of the form (27.7).)

The oscillations of the weight  $m$  make visible the transformation of kinetic energy into the potential energy, and vice versa. The theory of oscillations proves that in the process of a body's natural oscillations its total mechanical energy remains constant. Since in its extreme positions a body oscillating on a spring possesses only potential energy, its total excess energy while oscillating can be expressed by the formula (see Section 13-8)

$$E = kA^2/2 \quad (27.14)$$

where  $A$  is the amplitude of oscillations. In the equilibrium position the body possesses only kinetic energy. Therefore the total energy can be expressed also by the formula

$$E = mv_{\max}^2/2 \quad (27.15)$$

In other positions the total energy of an oscillating body is the sum of its potential and kinetic energies in this po-

sition:

$$E = U + K, \text{ or } E = kx^2/2 + mv_x^2/2 \quad (27.16)$$

where  $x$  is the displacement and  $v_x$  the velocity at point  $x$  where the energy is being measured.

### 27-13 Propagation of Oscillatory Motion in an Elastic Medium

When a fish bites, circles run over the water from the float. Water particles in contact with the float are displaced with it, and they involve in the motion other particles nearest to them, and so on.

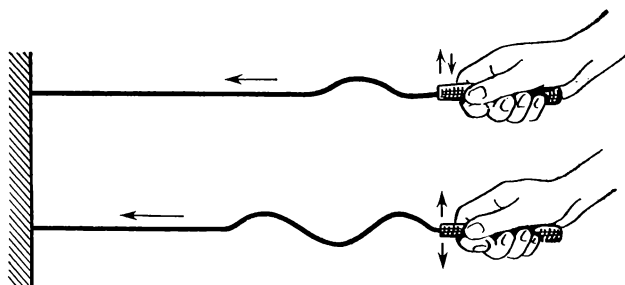


Fig. 27.13 Waves propagating along cord.

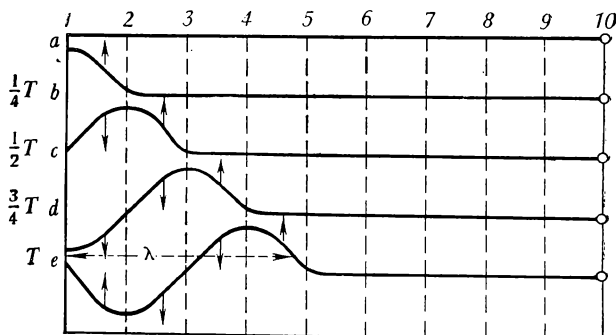


Fig. 27.14 Formation of wave along cord (arrows indicate direction of motion of points of wave).

The same happens to the particles of a tensioned rubber cord, one end of which is set in motion (Fig. 27.13). The term used for the propagation of oscillations in a medium is *wave motion*. Let us discuss in more detail how a wave along a cord is produced.

If we record the position of the cord every  $1/4 T$  after the vibrations at its initial point have started, we obtain the picture shown in Fig. 27.14b-e. The position  $a$  corre-

ponds to the start of vibrations at the initial point of the cord. Ten of its points are numbered, the dashed straight lines showing the positions of the same points of the cord at different moments of time.

At a time  $\frac{1}{4}T$  after it started vibrating, point 1 will occupy the uppermost position, point 2 just starting its vibrations. Since each successive point of the cord begins its motion after the preceding point, the position of the points in the interval 1-2 will be as shown in Fig. 27.14*b*. Still  $\frac{1}{4}T$  later point 1 will be in the equilibrium position and will move downwards, point 2 occupying the upper position (position *c*). At this instant point 3 starts moving.

In the time of a complete period the vibrations spread as far as point 5 of the cord (position *e*). After period  $T$  ends point 1 starts its second vibration. Simultaneously with it point 5 begins moving upwards in the course of its first vibration. From now on these points will vibrate in-phase. The totality of the points of the cord in the interval 1-5 forms a wave. As point 1 ends its second vibration, points 5-10 of the cord are involved in motion, that is, a second wave is formed.

If one observes the position of points having equal phases one notices how the phase appears to move from point to point to the right. Indeed, if point 1 has the phase  $\frac{1}{4}$  in position *b*, point 2 has the same phase in position *c*, and so on.

The waves in which the phase moves at a definite velocity are termed *travelling waves*. When one watches waves what one observes is the propagation of a phase, for instance of the crest of a wave. Note that all particles of the medium vibrate about their equilibrium positions and do not travel with the phase.

### 27-14 Energy Transport by Means of a Travelling Wave

We have learned above that in the process of sustained oscillations of some body, for instance a pendulum, its total energy remains constant, with the decrease in kinetic energy being accompanied by a simultaneous increase in potential energy, and vice versa. This is not so in the case of travelling waves.

The propagation of travelling waves involves the transport of energy from one vibrating particle to another. This is evident from the following example. When there is a splash on the surface of water caused, for instance, by a

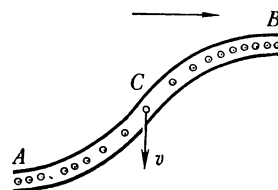
fish leaping out of it, waves spread out in circles from the location of the splash, transporting energy away from the point of their origin, with the surface of water calming down after the waves have run over it. To make the waves spread continuously, new energy would have to be continuously transmitted to the particles of water at the origin of the waves. For instance, by pulling periodically at a float one can produce a continuous set of waves on the surface of the water.

The transport of energy by a travelling wave has its explanation in the fact that the maxima of both kinetic and potential energies in such a wave correspond to the points of the wave which pass their equilibrium positions. We can demonstrate it with the example of a wave travelling along a cord.

Figure 27.15 depicts a part of a cord along which a wave travels to the right. It should be pointed out here that at rest the cord occupies a horizontal position. Hence, when a wave travels along the cord, it is not deformed around points  $A$  and  $B$ , maximum deformation of the cord being at point  $C$ . Therefore the maximum potential energy of elastic deformation of the cord coincides with point  $C$ , which passes the position of stable equilibrium.

But point  $C$  has moreover the greatest velocity  $v$  as compared with the other points of the cord, that is, it has the maximum kinetic energy. With point  $C$  moving downwards, at the next moment the half-way position will be occupied by the point nearest to it from the right. This point receives the maximum energy, to transmit it in turn to a more distant point, and so on. Hence, the energy is transported by the travelling wave with the velocity of phase propagation. Theory proves that the energy transported by an elastic wave is directly proportional to the density of the medium, to the square of the amplitude and to the square of the frequency of vibrations.

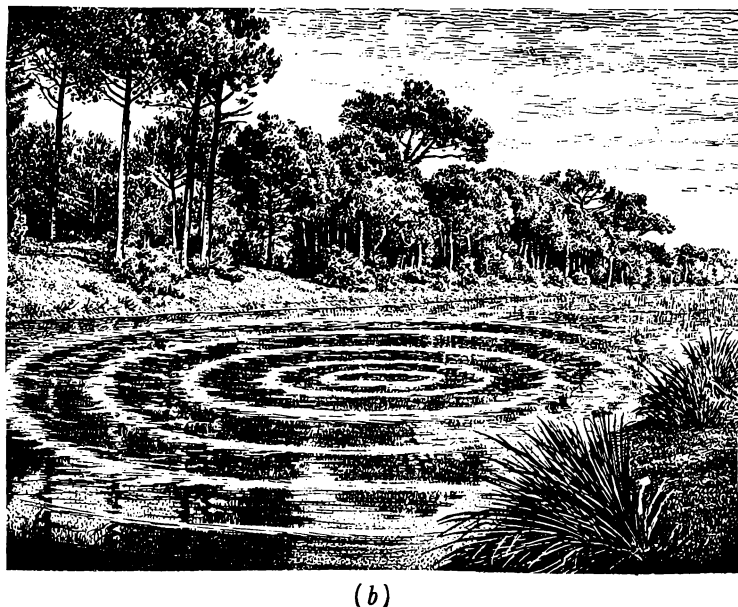
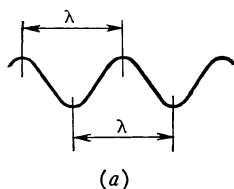
Fig. 27.15 Deformation of cord when wave propagates along it; points on cord indicate relative positions of particles in cord (arrow indicates direction of wave propagation).



## 27-15 Transverse and Longitudinal Waves

Let us return again to waves travelling along a cord as shown in Fig. 27.14, where the wave is seen to be travelling to the right and each point of the cord is moving upwards or downwards about its equilibrium position. The term used for a wave in which the particles of the medium move at right angles to the direction of propagation of the wave is *transverse wave*. Such waves consist of alternating crests and troughs.

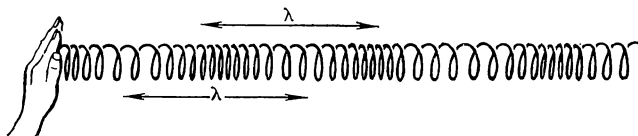
**Fig. 27.16** (a) Schematic representation of transverse wave  $\lambda$  is wavelength); (b) transverse waves on surface of water.



It was mentioned above that a cord along which a wave is travelling is subject to shear deformation causing a change in its shape. Transverse waves are only possible if the variation of shape is accompanied by the appearance of elastic restoring forces. Since this is an exclusive property of solids and of liquid surfaces, transverse waves can be excited only in solids and on the surfaces of liquids (Fig. 27.16).

Other types of waves also exist in nature. If one takes a long spring lying horizontally and gives it a push in the

**Fig. 27.17** Longitudinal wave in long spring.



axial direction as shown in Fig. 27.17, waves in the shape of alternating condensations and rarefactions will be excited in it. In this example the wave travels to the right, with each coil of the spring vibrating along its axis. The term used for a wave in which the particles vibrate along a straight line coinciding with the direction of wave propagation is

*longitudinal wave.* In such a wave the displacement of the particles takes place along a line joining their centres, that is, it results in variation of volume. Since restoring forces accompanying volume variation appear not only in solids and liquids but in gases as well, longitudinal waves are possible in solids, liquids and gases.

## 27-16 Waves and Rays. Wavelength

Figure 27.16*b* shows waves propagating along the surface of water. The bright circles are the crests of the waves, that is, the totality of points of maximum displacement from equilibrium positions. All these points vibrate in-phase.

For waves propagating not along the surface of a medium but inside it, the totality of points vibrating in-phase constitutes a surface of a definite shape. In an isotropic medium, that is, in a medium in which the phase velocity is independent of direction, the surfaces of equal phases of waves propagating from a point source are spheres. Such waves are termed *spherical*.

The term used for a continuous locus of points vibrating in-phase is *wave surface* (for instance, the bright circles in Fig. 27.16*b*). The leading surface, that is, the furthest from the source generating the waves, is termed *wave front*.

The line along which the wave front spreads is termed *ray*. It can be easily seen that in an isotropic medium the ray is always normal (perpendicular) to the wave surface. In an isotropic medium all rays are straight lines. In this case every straight line connecting the point of location of the source with an arbitrary point on the wave front is a ray.

In an isotropic medium the wave front travels with a constant velocity. Thus during one period of oscillations of the source responsible for the waves the wave front travels a strictly defined distance  $\lambda$ . Since each particle in the wave is involved in forced vibrations, the frequency of such vibrations is equal to the frequency of the source.

The term used for the quantity  $\lambda$  characterizing the displacement of the wave surface during one period as a function of the medium and the vibration frequency is *wavelength*. The measure of the wavelength is the distance travelled by a wave front during one period of oscillations of the wave source. Taking into account what was said in Section 27-13, we can define the wavelength also as the distance between two adjacent points vibrating in-phase, the



points being on the ray (see Fig. 27.14). (Prove that the distance between any two points of a travelling wave lying on the same ray and vibrating in-phase always contains an integral number of wavelengths or an even number of half-waves, and that the distance between any two points on a ray vibrating in opposite phases always contains an odd number of half-waves.)

For transverse waves (see Fig. 27.16a) the wavelength is equal to the shortest distance between two successive crests or troughs. For longitudinal waves (see Fig. 27.17) the wavelength is equal to the shortest distance between the centres of two adjacent condensations or rarefactions.

### 27-17 Velocity of Wave Propagation

We recall that wave propagation in a medium involves displacement of phase (see Section 27-13). The term used for the velocity of propagation of a characteristic point of a wave (e.g. of a crest) is *phase velocity*. Phase velocity is equal to the distance travelled by the phase of the wave divided by time. Since the phase of the wave travels a distance  $\lambda$  in time  $T$ , it follows that

$$v = \lambda/T \quad (27.17)$$

Since  $T = 1/f$ , we have

$$v = \lambda f \quad (27.18)$$

It is an established fact that phase velocity is determined by the physical properties of the medium. In a specific medium there is a unique correspondence between frequency and wavelength with waves of higher frequency having shorter wavelengths, as demanded by formula (27.18). This makes it possible to characterize waves in a medium not by their frequencies but by their wavelengths  $\lambda$ . At this stage it should be remembered that when a wave passes over from one medium into another, its frequency (and period) remain unchanged, but the wavelength  $\lambda$  changes in accordance with the change in the phase velocity  $v$ .

Note that a decrease in the phase velocity is accompanied by a proportional decrease in the wavelength. Hence the wavelength can be used as a characteristic of a wave only if it propagates in one medium.

## 27-18 Combination of Two Vibrations in Same Line

In practice we often witness superposition of vibrations. For instance, when two electric motors turning at different r.p.m. are mounted on the floor, the floor is subjected to a complex vibration resulting from the superposition of vibrations caused individually by each motor.

The resulting vibration might turn out to be quite intricate. A detailed study of such vibrations is beyond the scope of this book. However, to understand phenomena described below one should be familiar with the simplest cases of combination of vibrations. Consider the combination of two harmonic vibrations of the same frequency in the same line. The combination of such vibrations can conveniently be achieved by the graphic method.

If the graphs of displacements  $x'$  and  $x''$  versus time are available, we can find the resulting vibration by algebraic addition of the displacements  $x'$  and  $x''$  at every moment of time. (Figure 27.18 shows such addition performed for points  $B$  and  $C$ .) Hence, the resulting displacement at every point is expressed by the relation

$$x = x' + x'' \quad (27.19)$$

Connecting the ends of the ordinates  $x$  thus obtained by a smooth curve, we get the resulting vibration.

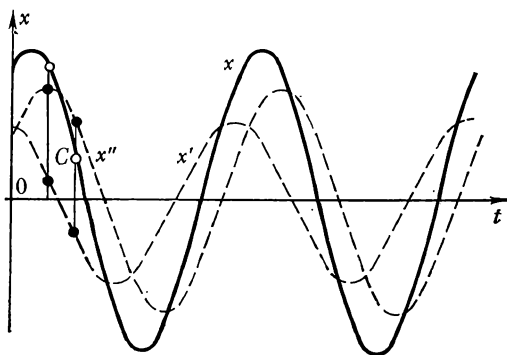


Fig. 27.18 Resulting oscillation (solid line) is at every point equal to algebraic sum of oscillations being added.

It can be seen from Fig. 27.18 that the combination of harmonic vibrations of the same frequency results in a harmonic vibration of the original frequency. The summation of such vibrations can be performed by a simpler method without resort to graphs.

Suppose the vibrations to be combined are described by the equations

$$x' = A' \sin(\varphi'_0 + 2\pi ft) \quad \text{and} \quad x'' = A'' \sin(\varphi''_0 + 2\pi ft)$$

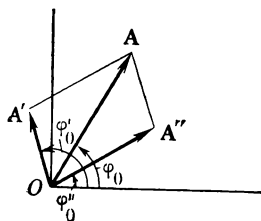
The theory of oscillations proves that the amplitude and the initial phase of the resulting vibration

$$x = A \sin(\varphi_0 + 2\pi ft) \quad (27.20)$$

can be found by vector summation of the amplitudes  $A'$  and  $A''$ . The method is as follows. Draw a horizontal straight line from an arbitrary point  $O$  (Fig. 27.19) from which the initial phases are to be measured. Draw from point  $O$  vectors  $A'$  and  $A''$  whose positions are determined by the initial phases  $\varphi'_0$  and  $\varphi''_0$ . The amplitude of the resulting vibration is the diagonal of the parallelogram built on the vectors  $A'$  and  $A''$ , the angle  $\varphi_0$  being equal to the initial phase of the resulting vibration. Figure 27.19 depicts the combination of vibrations whose graphs are shown in Fig. 27.18.

Note that  $A'$ ,  $A''$  and  $A$  are mobile radii of the vibrations being added. In Fig. 27.19 they are depicted at the initial moment. (Demonstrate that the projections of those vectors rotating counterclockwise at an angular speed  $\omega$  on the vertical axis represent the displacements of the respective vibrations.)

Fig. 27.19 Amplitude and initial phase of resulting oscillation are obtained by adding vectors  $A'$  and  $A''$  of two oscillations.



## 27-19 Reflection of Waves

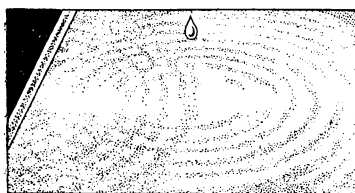
The reflection of waves can be observed in the following experiment. Place a ruler so that its edge is a little above the surface of the water (Fig. 27.20a) and excite waves on this surface by letting drops fall on it. The waves will be seen to reflect from the ruler's edge, the reflected waves travelling as if their origin was at point  $O_1$  symmetrical to point  $O$  with respect to the reflecting surface (the ruler).

Figure 27.20b is a schematic diagram of the reflection of the waves travelling from point  $O$  from the surface  $AB$ . Consider the reflection of waves at point  $A$ . The term for the ray  $OA$  is *incident*, and for the ray  $O_1A$  *reflected*. The angle  $i$  between the incident ray and the normal  $AD$  to the surface at the point of reflection is the *angle of incidence*, and the angle  $i'$  between the reflected ray and the same normal the *angle of reflection*.

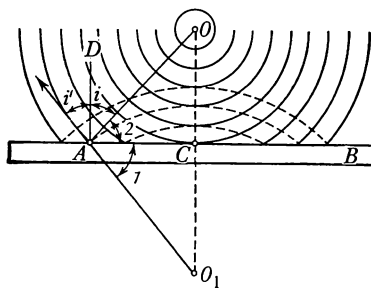
It can be seen from the diagram that the incident ray, the reflected ray and the normal  $AD$  at point  $A$  all lie in the same plane (in the plane of the diagram). Let us prove that the angles  $i$  and  $i'$  are equal. Because points  $O$  and  $O_1$  are

symmetrical ( $OC = O_1C$ ), the triangles  $OAC$  and  $O_1AC$  are equal and therefore  $\angle 1 = \angle 2$ . Since  $\angle i + \angle 2 = 90^\circ$ , it follows that  $\angle i' + \angle 1 = 90^\circ$ . This means that  $\angle i = \angle i'$ . Hence, the reflection of waves obeys the following two laws:

(1) The incident and reflected rays lie in the same plane with the normal to the reflecting surface at the point of incidence.



(a)



(b)

Fig. 27.20 Reflection of waves: (a) falling drop (the source of waves is seen in upper part); (b) schematic representation ( $i$  is angle of incidence,  $i'$  angle of reflection).

(2) The angle of reflection is equal to the angle of incidence

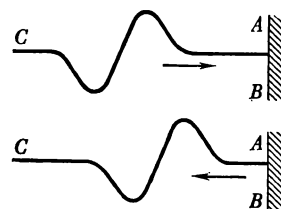
$$\angle i' = \angle i \quad (27.21)$$

Note that waves are not only reflected at the boundary separating two media, but generally also penetrate into the second medium. For instance, longitudinal waves may cross the boundary separating air and water in both directions.

Two typical cases of reflection can be observed in practice. Let there be a transverse wave travelling along a cord  $C$  with the crest leading the wave reflected from the surface  $AB$  (Fig. 27.21), the trough being in the lead. Such a case of reflection is said to be *reflection with the loss of a half-wave*, since the reflection involves a changeover to the opposite phase. Such reflection is typical for rigid reflecting surfaces, an example being the reflection of a wave travelling along a rubber cord from the fixed end of the cord.

Another case is reflection without change of phase. Such reflection is typical for compliant reflecting surfaces, an example being the reflection of a wave travelling along a cord from its free end.

Fig. 27.21 Reflection of wave with loss of half-wave.

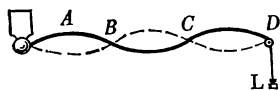


## 27-20 Standing Waves

The reflection of waves involves the superposition of vibrations. Reflected waves in Fig. 27.20a are seen to superimpose on the original waves. This means that every particle of water in places where the waves overlap takes part in a complex vibration.

If a small load  $L$  is tied to one end of a string running over a pulley  $D$  and the other end is tied to the hammer of an electric bell (see Fig. 27.22), the vibrations of the hammer will excite waves, which, after being reflected at point  $D$ , will travel back to the bell. The waves resulting from the superposition of the direct and reflected waves are termed *standing waves*, because the phases in it are stationary. In a standing wave points are clearly visible that do not take part in vibrations. Such points are termed *nodes* ( $B$ ,  $C$  and  $D$ ). Observing the vibrations of the string, we notice that all the points in the intervals between adjacent nodes vibrate in-phase but with different amplitudes. The term for points of maximum amplitude is *antinodes* (for example, point  $A$  in Fig. 27.22).

Fig. 27.22 Standing waves in string.



Continuing to observe the string, we see that the nodes and antinodes remain stationary. The distance between the adjacent nodes is equal to a half-wave and the phases in adjacent half-waves of the standing wave are always opposite. If at a moment of time the shape of the string is as shown by the solid line in Fig. 27.22, half a period later it will be as shown by the dashed line.

In standing waves there is no energy transport as in travelling waves: the sum of the potential and kinetic energies of point spaced at quarter-wave intervals remains constant although energy exchange takes place between the points. This is because the direct and the reflected waves transport energy in opposite directions.

Note that standing waves can be excited only if the distance from the source of vibrations to the obstacle that reflects the wave is equal to an integral number of quarter waves.

## 27-21 Interference of Waves

Imagine waves propagating along the surface of water from different point sources. These waves will superimpose in places of intersection. However, from the place of intersection each wave will travel as if it had not met the

other wave. This means that in a medium the propagation of waves from one source does not interfere with the propagation of other waves. If waves from different sources excite vibration at every point along a straight line, the superposition of such waves results in a displacement equal to the algebraic sum of displacements due to each individual wave. This is called *superposition of waves*.

When waves of different frequencies propagate in a medium, the vibrations of its particles are not harmonic, for at every point of the medium the phase difference between two vibrations changes continuously. Thus there are no regular oscillations.

The situation is different when the frequencies of the waves are identical. The term used for wave sources vibrating at identical frequencies and with the phase difference remaining constant in the process is *coherent sources*. The waves excited by such sources in a medium are also said to be *coherent*. The phase difference between the vibrations excited by the individual waves remains constant at every point of the medium.

The superposition of coherent waves propagating in a medium results in a stationary pattern of vibrations of particles of the medium, which shows the vibration amplitudes to be different at different points of the medium. The term for mutual intensification or suppression of vibrations in different points of a medium resulting from the superposition of coherent waves is *interference*. An example of an interference pattern is the standing wave pattern along a cord, since the direct and the reflected waves are coherent.

Note that the most convenient for observation are coherent sources, exciting harmonic vibrations in a medium. In this case the vibrations resulting from the superposition of the waves are harmonic and the relations deduced in Section 27-18 are used to describe them.

Specifically, if the phases of such waves at some point of the medium are opposite, the resulting amplitude will be equal to the difference of amplitudes of individual vibrations; and if the overlapping vibrations are in-phase, the resulting amplitude will be equal to the sum of their amplitudes. If the amplitudes are equal, in the former case the medium at the point will remain at rest and in the latter it will vibrate with double amplitude. Nodes and antinodes of a standing wave (see Fig. 27.22) are examples of such points. The interference of coherent waves on the surface of water is shown in Fig. 27.23. The lines along which the water remains at rest are also visible.

**Fig. 27.23** Interference of waves on water (white regions are where oscillations are suppressed).

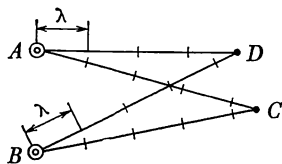


Figure 27.24 depicts two coherent in-phase sources  $A$  and  $B$  exciting waves of length  $\lambda$  in a medium. The term used for the distance in a homogeneous medium from the source to the specified point  $D$  is *wave path*. Note that in a homogeneous medium the wave path coincides with the geometrical path of the wave front from the source to the chosen point.

The following method is used to find the points of minimum (or maximum) amplitude resulting from interference. One has to find the difference in the wave paths from the sources to the specified point, that is  $BD - AD$ , and the number of half-wave lengths in this difference. If the resulting number of half-waves is odd, the waves from the sources reach point  $D$  in opposite phases and there is maximum suppression of vibrations at this point. If the number of half-waves is even (or zero), maximum intensification of vibrations will take place at this point.

At point  $D$  the difference in the wave paths is  $\lambda/2$ ; therefore there is maximum suppression of vibrations at this point. At point  $C$  the difference in wave paths,  $BC - AC$ , is zero. This means the overlapping vibrations are in-phase, that is, there is maximum intensification at this point.

**Fig. 27.24.** Waves reach point  $C$  in-phase—amplitude at this point grows, waves reach point  $D$  out of phase—amplitude decreases.



(Making use of Section 27-18 prove that the above method is correct.)

To conclude the section one more remark. The presence of interference patterns in a phenomenon is a certain proof of its wave nature.

## 27-22 Mechanical Resonance

We suspend three pendulums of different lengths from an elastic thread and add a fourth one with a length equal to the length of the second (Fig. 27.25).

If the second pendulum is made to swing with an amplitude  $A$ , all other pendulums will start oscillating, but the first and the third will soon stop after initial displacement, while the fourth will swing with an ever growing amplitude, finally becoming almost as great as  $A$ . The second pendulum will be observed to stop at that instant. Then the second pendulum starts swinging and continues to do so while the fourth stops, and so on. In this way the second and the fourth pendulums exchange energy via the thread  $BC$ .

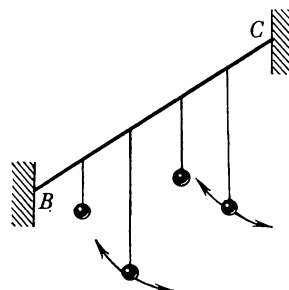
In this case alternating oscillations of the pendulums are caused by periodic jerks of the thread  $BC$  in time with the swings of the pendulums. The same method is used when we set a swing with someone on it in motion. The attenuation of the oscillations of the first and the third pendulums is due to the fact that they are out of time with the jerks of the thread.

The term used for an oscillating body exciting wave motion in the surrounding medium is the *vibrator*, and for the body whose amplitude of forced vibrations is at its maximum only at a definite frequency of the vibrator, the *resonator*. At the beginning of the experiment described above the second pendulum was the vibrator, the rest being resonators. Then the fourth pendulum became the vibrator, the second and the rest being resonators.

If various resonators are placed in a medium in the path of waves propagating in it, only the resonator whose natural frequency coincides with the wave frequency will oscillate with any great amplitude. This method can be used to determine the frequency of oscillations of particles in a wave. This is the principle of the frequency meter, used to measure the frequency of an alternating current.

Consider now the resonance phenomenon. Place a variable-frequency vibrator in a medium. Place a resonator some

Fig. 27.25 Resonance of pendulums (of second and fourth).





distance away from the vibrator and bring the vibrator into operation. As long as there is a substantial difference between the frequency of the vibrator and the natural frequency of the resonator, the latter will vibrate with a small amplitude. If the frequency of the vibrator is gradually brought closer to that of the resonator (say, from below), its amplitude will slowly rise. A sharp rise in the resonator's amplitude will take place only when the difference between the vibrator frequency and the natural frequency of the resonator become quite small. As soon as the vibrator's frequency begins to exceed the resonator frequency its amplitude drops sharply.

The term used for the frequency of the external force an increase or decrease in which results in a decrease in the amplitude of forced oscillations of a system is the *resonance frequency* of the system. The term *resonance* applies to the phenomenon of rapid rise in the amplitude of forced vibrations of a system as the frequency of an external force approaches its resonance frequency.

It was established above that the amplitude of oscillations of a system during resonance is greatly affected by the resistance of the medium (friction). The less the resistance the greater the amplitude of oscillations. With small resistance forces the oscillations at resonance might become so intense that they destroy the system. Note in addition that energy transported from the vibrator to the resonator is maximum at resonance. The experiment with the pendulums (see Fig. 27.25) is a good illustration.

Resonance is very important in nature and technology. Resonance is not an exclusive feature of mechanical processes. It is used in electrical engineering, in optics and in nuclear physics. It is the basic principle used in radios, television sets, etc.

Quite often resonance is harmful. For instance, the cabinet of a radio set sometimes rattles at particular frequencies, the foundations of rhythmically working machines wear out too soon, etc. In aviation resonance can cause a plane to break up in flight. This is the reason why new prototypes of aircraft are tested at various engine r.p.m.'s, at various flight speeds and in greatly varying conditions. During testflights the pilot observes resonance vibrations of individual parts of the plane so that they can be eliminated in the final stages of development. Theory and experiment combine in preventing resonance where it is harmful.

# Sound Waves and Ultrasonic Waves

# 28

## 28-1 What Is Sound!

Let us now look into the physical nature of sound phenomena.

A tuning fork is known as the instrument for the production of a pure tone. When a tuning fork emits sound, a ball brought into contact with its leg recoils from it because it vibrates (Fig. 28.1). Experience shows that sound is always produced by some vibrating body, which in the process of its vibrations excites mechanical waves in the surrounding medium (Fig. 28.2). When such waves reach the human ear, they excite forced oscillations of the eardrum and we hear sound. Mechanical waves that make a man sense sound are termed *sound*, or *audible, waves*.

Sound waves in air consist of compressions and rarefactions, that is, they are longitudinal waves. It is obvious that man can sense sound only if between the source of the sound and his ear there is a medium in which the sound waves can propagate. Sound cannot be transmitted in a vacuum. The following experiment proves this.

An electric bell is suspended by low-elasticity threads, connected to an ac power source and covered with a bell jar (Fig. 28.3). The sound of the ringing bell is still clearly heard. Next we pump the air out of the bell jar. As the air inside gets more and more rarefied the sound becomes weaker, and when the rarefaction is high enough the sound dies down completely.

In the course of the study of acoustic phenomena it was established that by no means all mechanical waves make a man sense sound. Only waves with frequencies lying in the range from 16 to 20 000 Hz were found to be audible.

Fig. 28.1 Ball bounces off tuning fork.

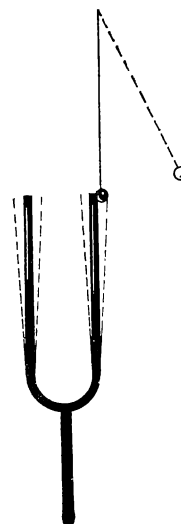


Fig. 28.2 Propagation of sound waves through air,

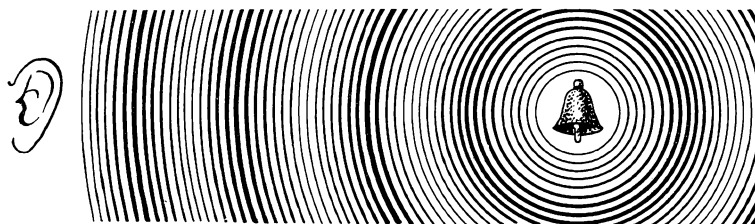
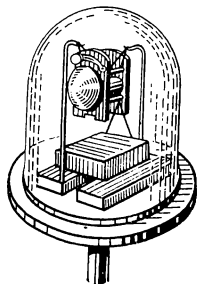


Fig. 28.3 In absence of air inside bell jar, ringing bell does not produce sound waves.



Note, however, that the individual upper and lower limits may vary to a certain extent.

Hence, a human being perceives sound only if the following four conditions are satisfied:

- (1) there is a source of sound;
- (2) there is an elastic medium between the ear and the sound source;
- (3) the frequency of the source ranges from 16 to 20 000 Hz;
- (4) the sound waves are powerful enough for the human to sense sound.

## 28-2 The Velocity of Sound

Since human beings sense sound mainly from waves propagating in air, we will now look at how the velocity of sound is measured, specifically in air.

We all know that in a thunderstorm we first see the lightning and then hear the thunder. The explanation is that the propagation velocity of light is several hundred thousand times greater than that of sound. Since the travelling time of light is very short, it can be neglected in measurements of sound velocity.

The experimental procedure for measuring sound velocity in air is as follows. Two experimenters take up positions at a specified distance from one another (about 1-2 km). One of them sends a light signal accompanied by a loud sound (for instance, a shot into the air) and the other starts a stop-watch the moment he sees the signal and stops it the moment he hears the sound. If one finds from the readings of the stop-watch the time the sound travelled, one can easily compute its velocity. Such experiments yield the result that the sound velocity in air at  $0^{\circ}\text{C}$  is 332 m/s, rising with the temperature.

Since the propagation velocity of waves depends on the type of medium and its state, sound velocity also depends on these factors. For instance, sound velocity in water is 1450 m/s and in steel it is 5000 m/s. (Explain why when pressing your ear to a rail you hear the train coming before you are able to hear it through the air.)

## 28-3 Loudness and Intensity of Sound

The sounds we hear produce quite different sensations. This difference is purely subjective, that is, different people are affected by sound in different ways. For instance, one

person may hear the same sound as a loud one, another may hear it as a soft one.

Hence, the *loudness*, or *sensation level*, of a sound is a subjective concept and any numerical estimate of it can only be a matter of convention. However, an objective measure of this quality of sound can be introduced, which can be equally applied to everyone. The term used for this objective measure of sound loudness is *sound intensity*. The measure of intensity  $J$  is the energy transported by sound waves in unit time across a unit cross-sectional area normal to the direction of wave propagation. Hence, the unit for measuring sound intensity in the SI system is

$$1 \text{ J}/(\text{m}^2 \cdot \text{s}) = 1 \text{ W}/\text{m}^2$$

We recall that the energy transported by the waves is directly proportional to the square of the amplitude and the square of the frequency (see Section 27-14). Therefore sound intensity, too, is proportional to the square of the amplitude and the square of the frequency of the sound wave.

In the case of spherical waves propagating from a source, sound intensity is inversely proportional to the distance from the source to the receiver. Indeed, if the source transmits an energy  $E$  to the waves in the time  $t$ , sound intensity  $J$  at a distance  $R$  from the source will be expressed by the formula

$$J = E/tA \quad (28.1)$$

Thus, for a spherical wave  $A = 4\pi R^2$  we obtain

$$J = \frac{E}{4\pi R^2 t}, \quad \text{or} \quad J = \frac{P}{4\pi R^2} \quad (28.2)$$

where  $P = E/t$  is the *power output* of the sound source.

Loudness is known to increase with amplitude and to decrease with distance. It has been established by experiment that variations in the amplitude of a sound wave affect only the volume of the sound and have no effect on other qualities of the sound.

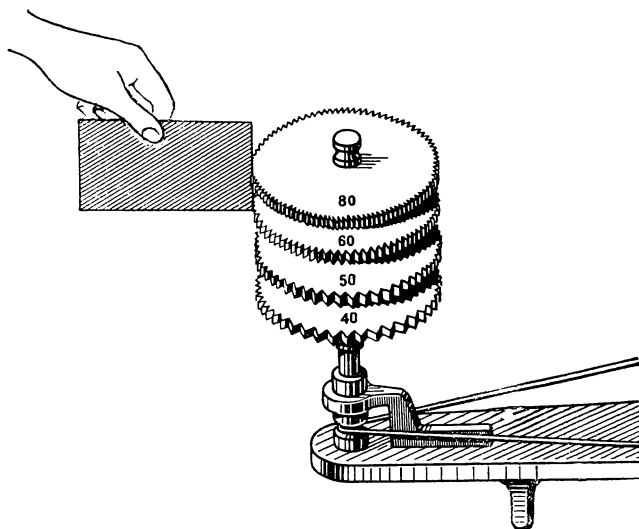
The human ear is very sensitive. The minimum intensity of sound waves able to produce a sensation of sound is termed the *threshold of audibility*. It depends on the frequency of oscillations. For instance, at a frequency of 2000 Hz the threshold is  $2 \times 10^{-12} \text{ J}/(\text{m}^2 \cdot \text{s})$ . At lower frequencies the threshold is much higher.

### 28-4 Pitch and Timbre of Sound

Another quality of sound a human can distinguish is *pitch*. For instance, it is easy to distinguish between the hum of a mosquito and the buzz of a bumblebee. The hum of a mosquito in flight is said to be high-pitched and the buzz of a bumblebee low-pitched.

Let us demonstrate by experiment that pitch is an objective characteristic of sound and is uniquely determined by the frequency of the sound wave. We rotate a system of toothed wheels of equal diameter but with a different number of teeth (Fig. 28.4). Pressing a small piece of cardboard

**Fig. 28.4** Pitch of sound produced by vibrating piece of cardboard depends on frequency at which it is struck by wheel's teeth.



against the teeth of each wheel in turn, we observe that the pitch of the sound rises with the number of teeth, that is, with the frequency of vibrations of the cardboard.

The term for sound of a strictly defined frequency is *tone*. The quality of sound dependent on its frequency is characterized by the pitch of the tone. It is conventionally assumed that a higher tone corresponds to a higher frequency of oscillations.

In some cases the pitch is characterized by the length of the sound waves in air (see Section 27-17). Indeed, from formula (27.18) for air at 0°C

$$\lambda = \frac{332 \text{ m/s}}{f} \quad (28.3)$$

It follows from this formula that the higher the tone the shorter its wavelength. In using wavelength to characterize the pitch one should keep in mind  $\lambda$ 's dependence on the type medium: in different media the wavelengths corresponding to the same tone are different. It can be easily reasoned that the wavelength will be greater in a medium with a higher sound velocity. Note that sometimes one hears complex sounds which cannot be resolved into individual tones without special instruments. Such sounds are heard as *noise*.

There is yet another property of sound in addition to loudness and pitch which a human ear can distinguish. The quality of sound that corresponds to the complexity of wave form produced by the source is its *timbre*, or *tone quality*. Timbre helps us to identify the person who is speaking or singing, or which instrument is playing, etc.

Some sources of sound can generate standing waves, for instance a vibrating string. When all its sections vibrate in phase, the string produces a definite tone termed *fundamental*. The string can also vibrate in higher modes, similar to those shown in Fig. 27.22, when two or more of its sections vibrate in opposite phases. The frequency of the tones produced by the string vibrating in a higher mode is always a multiple of the fundamental frequency; these tones are called *harmonic*.

Since the relative intensity of the harmonics produced by different types of sound sources is different, each one has a characteristic timbre distinguishable from the timbre of other sources, even if all sources produce the same fundamental frequency. This, of course, is not true of sound sources producing pure tones (i.e. only the fundamental frequency) of which the tuning fork and various types of electroacoustic generators are examples. Sources of pure tones are used to tune musical instruments and for many scientific applications.

### 28-5 Interference of Sound Waves. Beats

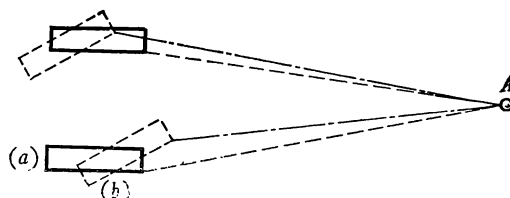
Interference of sound waves is the cause of a nonuniform distribution of the amplitude of the resulting oscillations at various points in space. It manifests itself either in the intensification or the suppression of sound at those points.

The two legs of a tuning fork produce coherent sound waves and this makes it possible to observe interference of sound waves at short distances from a tuning fork. As the tuning fork is rotated about its axis, the intensity of the sound changes (Fig. 28.5). Indeed, in the course of

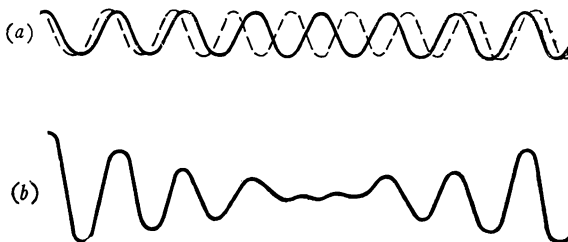
rotation the difference in wave paths registered at some point  $A$  will change continuously (see positions  $(a)$  and  $(b)$  in Fig. 28.5), that is, there will be an intensification in sound alternating with suppression. This is clearly heard when the fork is rotated.

At great distances from the fork, where the phase difference between waves radiated by the individual legs of the fork becomes small, there is practically only one wave.

**Fig. 28.5** Interference of sound waves accompanying rotation of vibrating tuning fork about its axis (fork is viewed from above; position  $(b)$  is shown by dashed line).



**Fig. 28.6** Production of beats:  $(a)$  waves from two sound sources with close periods of oscillations;  $(b)$  result of superposition of waves.



If two tuning forks of the same frequency are simultaneously excited, they will sound *in unison*. If a small piece of modeling clay is attached to one leg of a tuning fork, the forks will no longer sound in unison. In this case the forks will be heard to produce alternating sharp rises and falls in intensity of sound, unpleasant to the human ear and termed *beats*. The reason for the beats is that the phase difference between vibrations produced at a definite point by the waves coming from individual forks gradually changes from the in-phase to the out-of-phase (Fig. 28.6) and then again to in-phase, and so on.

It has been established that the frequency of beats is equal to the difference between the frequencies of the vibrations being added. Hence, the closer the frequencies the lower the beat frequency. This fact is exploited in tuning musical instruments. Making both the tuning fork and the string sound at the same time and listening for the beats, the tuner changes the tension of the particular string until it sounds in unison with the fork.

## 28-6 Reflection and Absorption of Sound

The reflection of sound waves from boundaries separating different media is of great practical importance. Consider an experiment which illustrates the laws of reflection of sound (see Section 27-19).

Put a wrist-watch on the bottom of a measuring glass. Choose a distance at which the watch is no longer heard, then place a glass plate above the measuring glass as shown in Fig. 28.7. The watch will then be heard. By changing the inclination of the plate and the position of the ear it can be verified that the angle of incidence is equal to the angle of reflection.

An interesting case of reflection occurs when the reflecting surface is at right angles to the direction of wave propagation. In this case the sound wave returns after reflection back to the source. The return of the sound wave after reflection to its source is termed *echo*.

It has been established that man continues to sense sound one-tenth of a second after the oscillations of his eardrums have stopped. This means that, if the distance from an ear to the reflecting surface is small, the echo will merge with the original sound, the only effect being a slight prolongation of the latter. Consequently, the echo can be separated from the original sound only if the distance between the ear and the reflecting surface is great enough.

Echo makes it possible to estimate the distance between the ear and the reflecting surface. Let the distance from the reflecting surface  $BC$  to the ear  $A$  be  $l$  (Fig. 28.8). If the interval between the time one hears the original sound **signal at point A and the reflected signal is  $t$  and the sound velocity is  $v$ , then  $2l = vt$ , or**

$$l = vt/2 \quad (28.4)$$

Obviously, the sound signal should be a short one for a long signal will merge with the echo and it could not be possible to measure  $t$ . (Demonstrate that for a sound velocity in air of 344 m/s (at 20°C) the echo will be heard separately from the original signal if the distance from the observer to the reflecting surface exceeds 17.2 m.)

In a closed room the sound is repeatedly reflected from the walls and is heard long after the source of sound ceased to function. The residual sound remaining in an enclosure is termed *reverberation*. For small enclosures the time of reverberation should be about one second. The time of reverberation greatly affects the quality of sound in concert halls, since an excessive time of reverberation makes it

Fig. 28.7 Reflection of sound waves from glass plate.

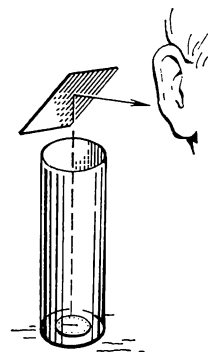
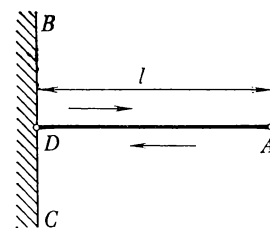


Fig. 28.8 Return of sound wave to point A after reflection at point D is used to find distance  $AD$ .



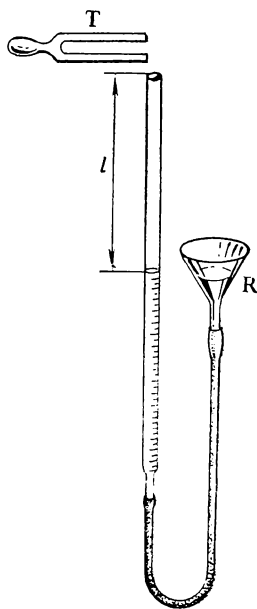


impossible to listen to music and a time of reverberation too short makes the tones sound flat and unfinished.

At the boundary between two media sound is also partially absorbed, that is, part of the energy of the sound waves is spent on overcoming the resistance of the vibrating particles of the second medium and turns into heat. That is why the walls of an enclosure affect its acoustics. For instance, a whitewashed wall absorbs eight per cent of the energy of sound waves and a carpet about twenty per cent. Accordingly, in a room full of furniture a tone sounds hushed and in an empty room the same tone sounds loud.

### 28-7 Acoustic Resonance

Fig. 28.9 Experiment demonstrating acoustic resonance.



Acoustic resonance can be obtained from two tuning forks of equal frequencies mounted on wooden boxes, called *resonators*, used to intensify the sound.

Arrange the forks one meter apart with the open ends of the boxes facing each other. Strike one of the forks by a rubber hammer—a loud tone will result. An instant later stop this fork—the tone will be less loud but will not die down altogether. The explanation is that sound waves excited the vibrations of the other fork tuned in resonance with the first. Indeed, if we were to stop the other fork, the tone would die down.

Note that in this experiment the columns of air enclosed in the boxes also take part in the resonance. The dimensions of the boxes are chosen so as to make the periods of natural vibrations of the air columns in them coincide with the periods of natural vibrations of the tuning forks.

There is another experiment which helps us observe the resonance of an air column. A tall vertical tube and a rubber tube with a reservoir R are connected (Fig. 28.9). Pour water into the reservoir and lift it level with the upper opening of the tall tube. Next place a vibrating tuning fork close to this opening and start lowering the funnel. Somewhere along the length of the air column, say at  $l$  centimetres from the top, there will be a sharp increase in the intensity of sound—resonance will set in. The resonance vanishes when the funnel is lowered still further.

### 28-8 Ultrasound and Its Applications

The human ear does not sense mechanical waves with a frequency higher than 20 000 Hz as sound. The term for such waves is *ultrasonic waves*, or *ultrasound*. Ultrasound

is strongly absorbed in gases and much less in solids and liquids. Because of that ultrasonic waves can travel long distances only in solids and liquids.

Since the energy transported by a wave across a unit area is proportional to the density of the medium and the square of the frequency, ultrasound can transport much more energy compared to sound waves. The short wavelengths of ultrasound also make it easier to design vibrator arrays for directional radiation and, hence, to concentrate energy. All this favours the wide use of ultrasound in technology.

In the *echo sounder*, used to measure the depth of the sea (Fig. 28.10), shorter wavelengths mean greater accuracy. The ship is equipped with a source and a receiver of ultrasound of a definite frequency. The source sends short pulses of ultrasonic waves and the receiver receives reflected pulses. Knowing the interval between the time the pulse was sent and received, one can, by using formula (28.4), compute the depth of the sea. A sonar, used to measure the distance to some obstacle in the way of the ship, acts in a similar way. In the absence of an obstacle the ultrasonic pulses do not return to the ship.

An interesting fact is that some animals, for instance bats, have organs operating on the sonar principle, enabling them to find their way in darkness. Dolphins, too, have a perfect sonar.

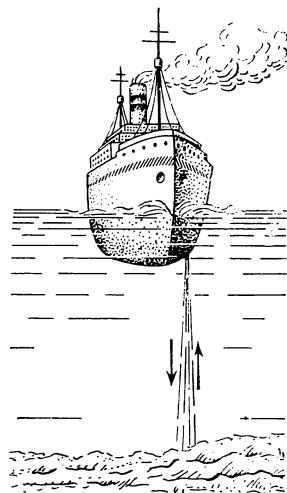
By passing ultrasonic waves of appropriate energy through a liquid its particles can be greatly accelerated so that they will act with a marked effect on various bodies immersed in the liquid. This principle is used to accelerate various technological processes (for instance, preparing solutions, cleaning components, tanning leather, etc.).

Intense ultrasonic vibrations in a liquid result in such great accelerations of its particles that instantaneous cavities are formed in the fluid whose collapse creates numerous shocks. We know that the term for this phenomenon is cavitation. In such conditions the liquid displays high crushing power, which is utilized in making suspensions consisting of small solid particles dispersed in a liquid and emulsions consisting of small droplets of one liquid dispersed in another.

Ultrasound is also used to detect flaws in metal parts. The uses of ultrasound in modern technology are so numerous that even simply listing them presents difficulties.

Note that the term for mechanical waves with a frequency below 16 Hz is *infrasonic waves*, or *infrasound*. They, too, do not produce a sense of sound. Infrasonic waves originate at sea during hurricanes and earthquakes. The velocity of

Fig. 28.10 Operation of echo sounder.



infrasound in water is much greater than that of a hurricane or of those gigantic waves, called *tsunami*, which form at the time of an earthquake. This helps some sea animals able to sense infrasound receive signals warning them of impending danger.

## 29

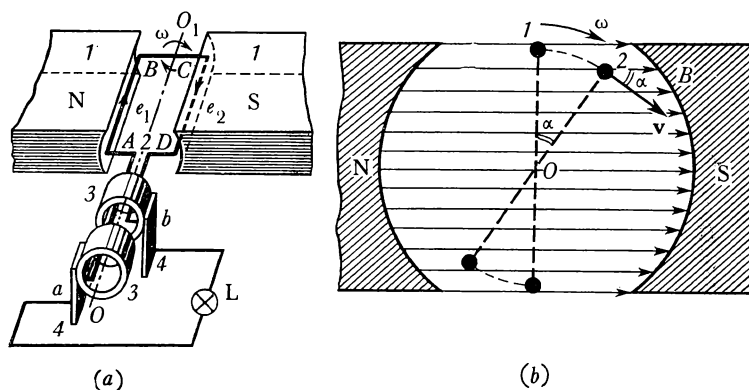
# Alternating-Current Circuits

### 29-1 Rotation of a Coil in a Homogeneous Magnetic Field

The emf and voltage in an ac main should vary in accordance with the harmonic law, that is, they should be sinusoidal (see Section 27-6). Deviations from this rule cause great power losses and for this reason it is strictly observed in practice.

Consider the production of a sinusoidal alternating current by a loop rectangular coil rotating at a constant speed in an homogeneous magnetic field. Let the ends of coil

Fig. 29.1 (a) Alternating-current induction generator; (b) angle between vectors  $\mathbf{B}$  and  $\mathbf{v}$  is equal to angle of rotation of coil from position 1.



$ABCD$  be connected to metal rings and the homogeneous magnetic field have an induction  $\mathbf{B}$  (Fig. 29.1a). Brushes  $a$  and  $b$  are in contact with the rings and are connected to the load  $L$ . If the coil is made to rotate clockwise about the axis  $OO_1$  with constant angular speed  $\omega$ , then emf's  $e_1$  and  $e_2$ , of equal magnitude and opposite direction, will appear in the segments  $AB$  and  $CD$ .

The wires  $AB$  and  $CD$  will move in a circle of diameter  $d = AD$  with a linear speed  $v = \omega d/2$ . If we start measur-

ing time and angle from the position of the coil  $I$  in Fig. 29.1b, the expression for the rotation angle of the loop  $\alpha$  will be

$$\alpha = \omega t, \text{ or } \alpha = (2\pi/T) t \quad (29.1)$$

where  $T$  is the time the coil takes to make a complete revolution. Since angle  $\alpha$  is equal to the angle between  $\mathbf{B}$  and  $\mathbf{v}$ , for the emf induced in segment  $AB$  (or  $CD$ , for that matter) we have the formula (see Section 26-3)

$$e_1 = Bvl \sin \alpha$$

where  $l$  is the length of conductor  $AB$  or  $CD$ . Note that the term for such conductors is *active conductors*, for it is only in them that the emf's are induced when the coil rotates in the magnetic field. The total emf in the coil is

$$e = 2e_1 = 2vlB \sin \alpha, \text{ or } e = 2(\omega d/2) lB \sin \omega t \quad (29.2)$$

Since  $\omega$ ,  $d$ ,  $l$  and  $B$  are constant, their product can be denoted by  $\mathcal{E}_{max} = \omega dlB$ . Then

$$e = \mathcal{E}_{max} \sin \omega t, \text{ or } e = \mathcal{E}_{max} \sin (2\pi/T) t \quad (29.3)$$

We recall that the maximum value of a sine is unity. Therefore  $\mathcal{E}_{max}$  in formula (29.3) has the maximum emf produced in the coil in the course of its rotation. The term used for  $\mathcal{E}_{max}$  is the *amplitude of emf*. Figure 29.2 shows the curve for a sinusoidal emf. Note that the common practice is to denote the instantaneous values of alternating current parameters by small letters and to reserve the capital letters for their amplitudes. For instance, the notation used for the instantaneous current is  $i$  and for its amplitude  $I$ . The notations for the voltages are  $u$  and  $U$  respectively.

In the above example the angular frequency of the alternating current  $\omega$  in formulae (29.2) and (29.3) coincides with the angular speed of rotation of the coil in the magnetic field, and the period  $T$  of the alternating current coincides with the period of rotation of the coil. Making use of the *frequency* of alternating current,  $f = 1/T$ , we rewrite formula (29.3) in the form

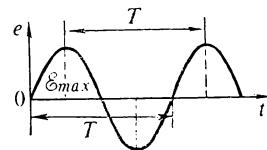
$$e = \mathcal{E}_{max} \sin 2\pi ft \quad (29.3a)$$

Denoting the number of revolutions of the coil per minute (r.p.m.) by  $n$ , we obtain

$$f = n/60 \quad (29.4)$$

The standard frequency of alternating current in the USSR is 50 Hz. This means that the emf and current in an

Fig. 29.2 Induced emf versus time curve for ac induction generator ( $T$ —period of alternating current).



ac circuit reverse their direction 100 times per second. Alternating currents of frequencies from about 20 Hz to about 100 000 Hz (i.e. 100 kHz) are termed *low-frequency currents*, and those of frequencies from several megahertz and higher are termed high-frequency currents (see Section 30-3).

## 29-2 The Induction Generator

Fig. 29.3 Distribution of induction  $\mathbf{B}$  on armature's surface.

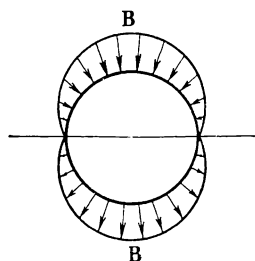
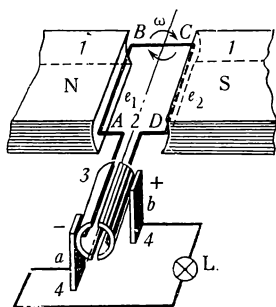


Fig. 29.4 Direct-current induction generator.



Electric machines using electromagnetic induction to transform mechanical energy into electric energy are termed *induction generators*.

The principal elements of an *ac induction generator* are shown in Fig. 29.1a: 1 is the *inductor* which sets up the magnetic field, 2 is the *armature* (the conductor in which the emf is induced), 3 is the metal *collector*, or *slip*, rings and 4 are *brushes* providing contact between the stationary and rotating conductors.

Since only relative motion of the conductor and the magnetic field is important for the generation of an emf, in practice the inductor is made to rotate. The term for it in this case is the *rotor*. The armature is made stationary and the term for it is the *stator*. The advantages of such design are that the rotor is an electromagnet with relatively small currents, obtaining them via sliding contacts (which operate reliably only at small currents), and that the load is connected to the generator via stationary contacts.

The rotor and the stator are made of steel with only a small gap between them; because of this vector  $\mathbf{B}$  in the gap is everywhere perpendicular to the stator's surface. Therefore vector  $\mathbf{B}$  is permanently perpendicular to the vector of linear velocity of the points on the rotor's surface, that is, to the vector of relative velocity of the magnetic field and the conductors in the armature. This means that the magnitude of angle  $\alpha$  in the expression for the emf,  $e_1 = Bvl \sin \alpha$  is always equal to  $\pi/2$ . Therefore, in order to induce in the conductors a sinusoidal emf, the magnetic poles are of a special shape which provides for a sinusoidal variation of vector  $\mathbf{B}$  along the rotor's circumference (Fig. 29.3).

When the rotor has a single pair of magnetic poles, the frequency of the rotor's rotation coincides with the frequency of the alternating current. When the number of pole pairs is two, the frequency of magnetic field variation is twice that of the rotor's rotation frequency, and a rotor of this type has to make 25 r.p.s. instead of 50 r.p.s. to produce

a current of the standard 50 Hz frequency. One pair of poles is the feature of turbogenerators whose rotors are driven by steam, while several pairs of poles are used in low-speed generators installed in hydroelectric power stations.

A schematic diagram of a *dc induction generator* is presented in Fig. 29.4. The only difference in design from that of the ac generator (Fig. 29.1) is in the device called *commutator*.

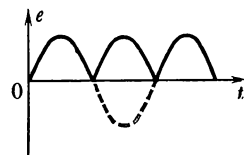
The commutator of a dc generator (3 in Fig. 29.4) is a split ring, each segment of which is connected to an end of the armature loop. The main purpose of the collector is to produce current flowing in one direction. This is accomplished by providing a constant contact between the left brush and the rising edge of the loop and between the right-hand brush and the falling edge. The emf versus time curve for such a generator is shown in Fig. 29.5. The dashed line shows, for the sake of comparison, the emf versus time curve for a solid ring.

Thus, the split-ring commutator rectifies the alternating current induced in the armature. To secure a more constant voltage from a dc generator, in practice the armature winding is divided into sections, each connected to a different pair of commutator segments.

When a generator is working, Ampere force acts on the conductors in its armature (see Section 25-9), opposing its rotation. This force increases in magnitude with the current flowing in the armature winding. Therefore, when the consumption of current produced by the generator increases, so does the power needed to rotate its armature. This is equally true of ac generators.

Note in addition that dc electric machines possess the property of convertibility, that is, they can work both as generators and as electric motors.

Fig. 29.5 Induced emf versus time curve for dc induction generator.



### 29-3 Effective Values of EMF, Voltage and Current

For a sinusoidal alternating current the values of the voltage and the current averaged over a period are zero and cannot therefore serve as characteristics of the current. However, the square of the current averaged over a period is not zero. Accordingly, if an instrument whose pointer is deflected in proportion to the square of the current is connected into an ac circuit, the pointer will stop at a specific point on the instrument's scale. What is the meaning of this reading?

Recall that the heat liberated in a conductor is proportional to the square of the current. Imagine that a thermal ammeter, whose action is based on the heating effect of electric current, is connected into a circuit. A definite reading on such an ammeter can be obtained by passing through it both a dc and an ac current. This makes it possible to introduce the concept of the *effective value* of an alternating current. The term effective value of an alternating current applies to a direct current which during one period of the ac current generates heat in a resistor equal to the heat generated by the alternating current in the same time interval in the same resistor.

Alternating-current ammeters are graduated in effective values of the current. Courses in electrical engineering contain the proof that the effective value of a sinusoidal alternating current is  $\sqrt{2}$  times less than its *amplitude*, or *peak*, value  $I_{max}$ , that is

$$I_{\text{eff}} = \frac{I_{max}}{\sqrt{2}} \approx 0.707 I_{max} \quad (29.5)$$

The definition of effective ac voltage is similar to that of effective ac current: effective ac voltage is the dc voltage across a resistor that causes current to flow in it which generates the same amount of heat per period as that generated in it per period by ac voltage. The relation between the effective and peak ac voltages is the same as between the currents:

$$U_{\text{eff}} = \frac{U_{max}}{\sqrt{2}} \approx 0.707 U_{max} \quad (29.6)$$

Similarly, for ac emf's the effective value  $\mathcal{E}_{\text{eff}}$  is  $\sqrt{2}$  less than its peak value  $\mathcal{E}_{max}$ :

$$\mathcal{E}_{\text{eff}} = \frac{\mathcal{E}_{max}}{\sqrt{2}} \approx 0.707 \mathcal{E}_{max} \quad (29.7)$$

Alternating-current voltmeters are usually graduated in effective values of ac voltage.

#### 29-4 Inductance and Capacitance in an AC Circuit

In an ac circuit there are, in general, phase differences between current, voltage and emf. Accepting the initial phase of the current as zero, we can write for the instant-

neous values of current, voltage and emf

$$i = I_{max} \sin \omega t \quad (29.8)$$

$$u = U_{max} \sin (\varphi + \omega t) \quad (29.9)$$

$$e = \mathcal{E}_{max} \sin (\psi + \omega t) \quad (29.10)$$

The parameter of the circuit responsible for irreversible losses of electric power converted into heat is termed *resistance*. At low frequencies the resistance of a circuit is independent of the frequency and equal to its dc value obtained from formula (18.18):

$$R = \frac{\rho_0 l}{A} (1 + \alpha t^\circ \text{C})$$

where  $\alpha$  is the temperature coefficient of resistance and  $t^\circ \text{C}$  the temperature.

The phase shift between the current and the voltage in an ac circuit containing only resistance, for instance, incandescent lamps, heaters, etc., is zero. This means that the current and the voltage in such a circuit are in-phase and that the electric energy is completely converted into heat. Connection of a coil with an inductance  $L$  into an ac circuit has the effect of reducing the current flowing in the circuit. The explanation is that an ac current flowing in the coil induces a back emf responsible for the effect. The parameter determining the magnitude of the above effect is termed *inductive reactance*.

Since the self-induced emf increases with the inductance and with the rate of current variation, the inductive reactance is directly proportional to  $L$  and to the ac angular frequency  $\omega$ :

$$X_L = \omega L \quad (29.11)$$

The following experiment, depicted in Fig. 29.6, clearly demonstrates the effect of inductive reactance on the current flowing in a circuit. Inserting the ferromagnetic core into the coil, we see that the lamp goes out; the withdrawal of the core makes it burn again. The explanation is that inserting a core greatly increases the coil's inductance. It should be noted that the voltage across an inductive reactance is ahead of the current flowing in it.

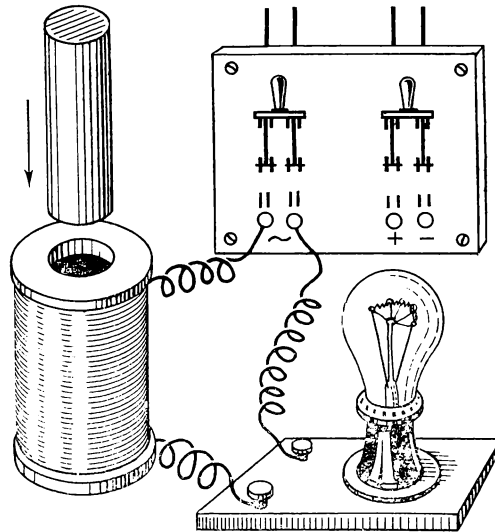
Direct current cannot pass through a capacitor because there is a dielectric between its plates. If a capacitor is connected into a dc circuit, the current in the circuit ceases as soon as the capacitor is charged.

Now let us connect a capacitor into an ac circuit. The capacitor's charge ( $q = Cu$ ) changes continuously because



of the changing voltage, and this makes ac current flow in the circuit. The current will be the greater the greater the capacitance and the greater the recharging rate, the frequency of the ac current.

Fig. 29.6 Ferromagnetic core inserted into coil increases inductive reactance and lamp stops burning; when core is withdrawn, lamp lights up again.



The resistance due to a capacitance in an ac circuit is termed *capacitive reactance*  $X_C$ . It is inversely proportional to the capacitance and the angular frequency  $\omega$ :

$$X_C = 1/\omega C \quad (29.12)$$

Comparing formulae (29.11) and (29.12) we see that the reactance of a coil is great at high frequencies and low at low frequencies, the opposite being true of capacitors. The voltage across a capacitive reactance  $X_C$  lags behind the ac current flowing in it.

The theory of alternating currents proves that when inductive and capacitive reactances are connected in-series, the combined reactance is equal to their difference:

$$X = X_L - X_C \quad (29.13)$$

Its character is inductive if  $X_L > X_C$  and capacitive if  $X_L < X_C$ .

Note in conclusion that the *average power* dissipated in a specific section of the circuit is expressed by the formula

$$P = IU \cos \varphi \quad (29.14)$$

The power spent on heating the circuit's resistance is

$$P = I^2 R \quad (29.15)$$

and gives the *actual* power dissipation.

It follows from (29.14) that in order to increase the power dissipated in the circuit its  $\cos \varphi$  should be made higher. (Explain why the maximum value of  $\cos \varphi$  is attained when  $X_L = X_C$ .)

### 29-5 The Transformer

One important advantage of ac over dc is that ac voltage can be easily varied with the aid of electromagnetic induction, whereas complex methods have to be used to vary dc voltage.

The device used for the transformation of ac voltage and current is termed a *transformer* (Fig. 29.7). It was invented

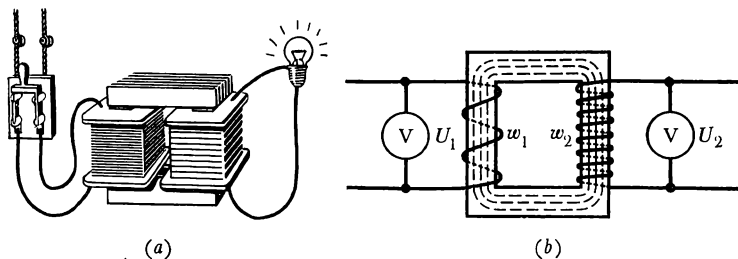


Fig. 29.7 (a) Transformer; (b) schematic representation of transformer's operation (dotted lines are lines of magnetic induction).

by the Russian scientist P. N. Yablochkov (1847-1894) in 1876. A transformer consists of a closed core made of soft magnetic steel or ferrite with two insulated coils called *windings*. The *primary* is connected to an ac main and the *secondary* to the consumer. The current in the primary establishes an alternating magnetic flux in the core which induces an equal emf in every turn of both windings. If the number of turns in the primary is  $w_1$  and in the secondary  $w_2$ , the emf induced in the windings will be directly proportional to the number of turns in each of them

$$\mathcal{E}_1 / \mathcal{E}_2 = w_1 / w_2 \quad (29.16)$$

When the secondary is disconnected from the consumer (open-circuit conditions for the transformer), the voltage  $U_2$  across its terminals is equal to  $\mathcal{E}_2$ . A small current  $I_0$  flows in the primary, called *open-circuit current*. Since the voltage drop on the resistance of the primary is very

small, the voltage  $U_1$  is quite close to  $\mathcal{E}_1$ ; we can say that  $U_1 = \mathcal{E}_1$ .

Hence, in an open-circuit condition the voltages across transformer windings are directly proportional to the number of turns in each of them:

$$U_1/U_2 = w_1/w_2 \quad (29.17)$$

If the number of turns in the secondary,  $w_2$ , is greater than in the primary,  $w_1$ , the transformer is called a *step-up transformer*; if  $w_2$  is less than  $w_1$ , it is a *step-down transformer*. The ratio of the number of turns in the primary to that in the secondary is termed *transformation ratio*,  $n$ :

$$n = w_1/w_2 \quad (29.18)$$

Hence, the transformation ratio of a step-down transformer exceeds unity and of a step-up transformer is less than unity.

With the secondary circuit closed (a load connected across the secondary) the current  $I_2$  flowing in it sets up a magnetic flux in the core directed against the flux set up by the primary. The decrease in the flux in the core reduces  $\mathcal{E}_1$  in the primary and this causes the current in it to rise to a new value  $I_1$ , at which the increased magnetic flux compensates the opposite flux of the secondary so that the resulting flux in the core reaches its original value.

Since the magnetic flux of a coil is proportional to the number of turns in it and to the current, one can assume that approximately  $I_1 w_1 = I_2 w_2$  (actually  $I_1 w_1$  is somewhat greater than  $I_2 w_2$ ). Hence

$$I_1/I_2 = w_2/w_1 \quad (29.19)$$

that is, the current in the windings is inversely proportional to the number of turns in them.

The voltage drops across the resistances of the coils are not large. Therefore it may be assumed, as a good approximation that  $U_1 \approx \mathcal{E}_1$  and  $U_2 \approx \mathcal{E}_2$ , that is, that the expression (29.17) holds also for a loaded transformer. It follows from (29.17) and (29.19) that  $I_1 U_1 \approx I_2 U_2$ . Because of close coupling between the windings the values of  $\cos \varphi$  in both windings are approximately equal. Therefore the power  $P_1$  consumed by the primary from the ac main is almost equal to the power  $P_2$  transmitted by the secondary to the consumer and the ratio  $P_2/P_1$ , termed the *efficiency of the transformer*, is close to unity.

### 29-6 Induction Coil

The *induction coil* is a device used to transform low-voltage dc into high-voltage ac; it is a transformer of original design (Fig. 29.8).

When the switch  $S$  is closed, the current from the battery  $B$  passes through the adjustment screw  $A$ , the interrupter  $I$  on a steel spring, the primary coil  $P$  with a ferromagnetic core and returns to the battery. The current magnetizes

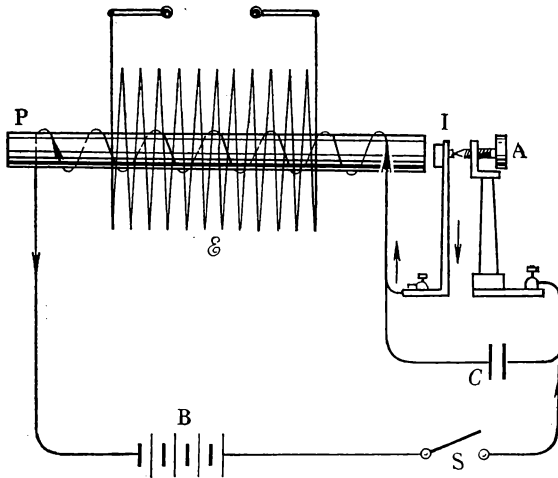


Fig. 29.8 Induction coil.

the core, the core attracts the interrupter and this breaks the circuit. Without the current the core is demagnetized and the spring returns the interrupter, closing the circuit. The whole process is then repeated. In this way a rapidly changing magnetic field is set up around the primary which induces an emf,  $\mathcal{E}$ , in the secondary. The terminals of the secondary winding, which has a great number of turns, are shown in the upper part of the figure.

It follows from (26.7) that self-induced emf is proportional to the rate of current variation. In the case of the induction coil this rate is at its maximum at the time the contact is broken. The voltage across the primary at the moment the contact is broken is so high that a spark appears, short-circuiting the primary and reducing the voltage across it (because the resistance across the contact gap drops as the spark develops). To prevent spark discharge a capacitor  $C$  is connected across the gap to remove the first voltage peak. Such a coil produces voltages across the secondary

up to 20 000 V. A modification of it, the ignition coil with the interrupter driven by the engine, is widely used as an ignition device for gasoline piston combustion engines.

### **29-7 Production, Transport and Distribution of Electric Energy**

It is a well-known fact that economic progress is based on the growth in power production. Since electric power stations are the main suppliers of industrial power, in the USSR great attention is being paid to the construction of new power plants and to increasing the power of those already in operation. Total power output of all electric power plants in the USSR in 1974 exceeded  $200 \times 10^6$  kW, and in 1980 this output will grow another  $100 \times 10^6$  kW.

The Soviet Union also has enormous reserves of hydroelectric power. Such giants as the Bratsk hydroelectric station, with a power output of  $4.5 \times 10^6$  kW, and, near Krasnoyarsk, the largest hydroelectric station in the world ( $6.0 \times 10^6$  kW), equipped with hydroelectric turbo-generators  $0.5 \times 10^6$  kW each, have already been built. A system of hydroelectric plants on the Angara river with a total power output of  $(12-15) \times 10^6$  kW is under construction.

At present the bulk of electric energy in the Soviet Union is being generated at thermal electric power plants. The power output of the largest electric power station in the world, at Krivoi Rog, is  $3 \times 10^6$  kW. Even more powerful stations, each with a power output of  $(4-5) \times 10^6$  kW, are being built. They will be equipped with turbogenerators up to  $0.8 \times 10^6$  kW each. Still more powerful turbogenerators of  $1.2 \times 10^6$  kW are being developed.

However, the quickest progress is in the construction of atomic power plants (see Section 42-4), which in the near future will take over first place in the production of electric energy.

Other methods of producing electric energy are also being used—solar, geothermal, wind, etc. Electric power plants using these sources of energy are being built. Plans have been laid to exploit tidal energy in the near future. For instance, a powerful tidal electric station to be built on the White Sea is presently being designed.

Electric energy can be used effectively only if it can be transported over great distances with minimum losses. To

this end the energy should be transported at high voltages. Transmission lines with voltages of 500 and 750 kV are already in operation; lines with voltages of over a million volts are at the blue-print stage.

Electric power stations in different parts of the Soviet Union are interconnected by means of high-tension transmission lines forming integrated power supply systems. Such systems have been devised for Siberia, Central Asia and the European part of the USSR. The integration of these systems will mean the creation of an electric power supply system for the whole of the USSR. Owing to a great difference in time between the Eastern and the Western regions of the USSR this will save up to  $40 \times 10^6$  kW in

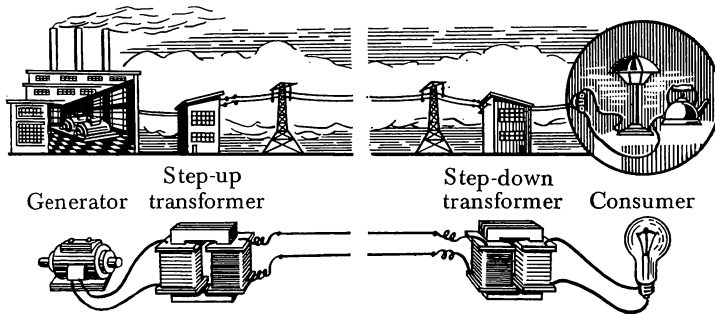


Fig. 29.9 Schematic representation of long-distance power transmission line.

electric power, that is, instead of building electric power stations of such output the power can simply be switched over to the zone where its consumption at that moment is highest. The integration of the power supplies of neighbouring socialist countries will also bring great benefits.

A simplified diagram of long-distance electric power transmission is presented in Fig. 29.9. In high-tension transmission lines direct current has an important advantage over alternating current: for the same voltage and the same lay-out direct-current lines are less liable to become sources of electric discharge in the atmosphere, which, of course, results in additional power losses. Direct-current transmission lines for a voltage of  $1.5 \times 10^6$  V are presently being designed.

## 30

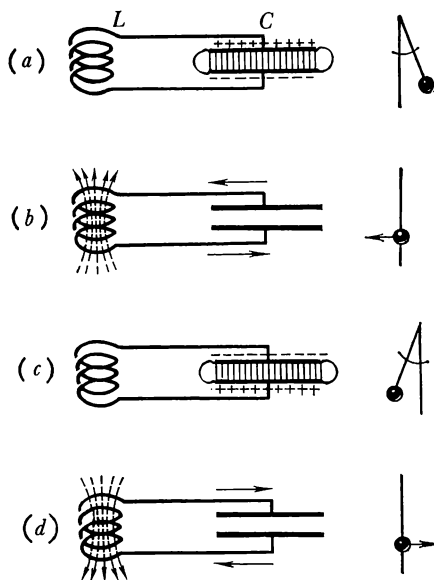
# Electrical Oscillations and Electromagnetic Waves

## 30-1 Transformation of Energy in a Closed Oscillatory Circuit

To produce electrical oscillations a circuit is needed in which the energy of the electric field can be transformed into the energy of the magnetic field, and vice versa. The term for such a circuit is *oscillatory circuit*.

Since a magnetic field is produced in a coil and an electric field in a capacitor, the simplest oscillatory circuit consists

**Fig. 30.1** Electrical oscillations in circuit consisting of capacitance and inductance; diagrams (a), (b), (c) and (d) depict successive instants at intervals of one quarter of period (for comparison pendulum oscillations are shown at right).



of a coil of inductance  $L$  and a capacitor of capacitance  $C$ . The resistance of the wires of which the circuit is made should be low, otherwise the circuit will not oscillate.

Consider the process of the generation of electrical oscillations in more detail. Charge the capacitor (with capacitance  $C$ ) so that the voltage across it is  $U_{max}$  and connect it to a coil (with inductance  $L$ ) (Fig. 30.1). Figure 30.1a depicts the moment when the capacitor's discharge has just started. At this moment there is an electric field in

the capacitor and no magnetic field in the coil. Therefore the entire excess energy is in the form of electric energy and is equal to  $E_{el} = CU_{max}^2/2$ .

As the charges rush from the capacitor to the coil, a self-induced emf is set up in it which opposes the increase in the current but is unable to stop it altogether (see Section 26-9). The current continues to increase until the capacitor is completely discharged. At this moment (Fig. 30.1*b*) the current in the circuit is at its maximum value  $I_{max}$ , and all the excess energy of the circuit is transformed into the energy of the magnetic field of the coil expressed by the formula  $E_{mag} = LI_{max}^2/2$ .

If the resistance of the circuit  $R$  is so small that thermal losses in the circuit can be neglected,  $E_{mag}$  will be equal to  $E_{el}$ . If  $R = 0$ , that is, in the case of natural oscillations in the circuit, then

$$\frac{CU_{max}^2}{2} = \frac{LI_{max}^2}{2} = \text{constant} \quad (30.1)$$

At the next moment the magnetic field in the coil starts to diminish and a self-induced emf sets up in it in support of the original current, causing the capacitor to recharge, that is, a transformation of magnetic energy into electric energy takes place. As soon as the magnetic field in the coil disappears, the capacitor starts discharging again (Fig. 30.1*c*) and a current of opposite direction continues to flow in the circuit until all the electric energy is again transformed into magnetic (Fig. 30.1*d*). After that the self-induced emf in the coil again recharges the capacitor and the circuit returns to the state shown in Fig. 30.1*a*. This completes the cycle in the circuit and subsequently everything is repeated in the same order.

One is bound to notice the similarity between the electric oscillations in a circuit and mechanical oscillations: the electric energy of a charged capacitor may be likened to the potential energy of a pendulum, and the magnetic energy of the current flowing in a coil to the kinetic energy of the pendulum (see Fig. 30.1).

The time of one complete oscillation is the *period* of electrical oscillations,  $T$ , and the number of oscillations per unit time is the frequency  $f$ , or  $f = 1/T$ .

Theory proves that in an ideal circuit ( $R = 0$ ) the period of oscillations, that is, the *period of natural oscillations*, is determined from the condition of equality of the reactances of coil and capacitor:

$$X_L = X_C, \quad \text{or} \quad L\omega_r = \frac{1}{\omega_r C} \quad (30.2)$$



The term for the angular frequency at which the above equality is satisfied is *angular resonance frequency* (of the ideal circuit), because when an external emf acts in the circuit it is at that frequency that the amplitude of its forced oscillations is at its maximum (see Section 27-22). (Explain why in the case  $R = 0$  the amplitude of oscillations must be infinite.)

It follows from (30.2) that

$$\omega_r = \frac{1}{\sqrt{LC}} \quad (30.3)$$

Since  $\omega_r = 2\pi/T$  (see Section 27-6) we obtain for the period of natural oscillations

$$T = 2\pi \sqrt{LC} \quad (30.4)$$

The relation (30.4) is the *Thomson formula*.

The formula for the *frequency* of natural oscillations is obviously

$$f = \frac{1}{2\pi \sqrt{LC}} \quad (30.5)$$

It follows from (30.5) that reducing  $L$  and  $C$  increases the frequency of oscillations. Maximum frequency in a *lumped circuit*, consisting of a separate capacitor and a separate coil, is limited by the capacitance and inductance of wires connecting them and is of the order of 500 MHz ( $500 \times 10^6$  Hz).

### 30-2 The Electron Tube Oscillator

Since every circuit of the type described in the previous section has a resistance  $R$ , the energy of the charged capacitor connected to the coil decreases with each oscillation because it is spent on heating the resistance. This means that in practice free oscillations in a circuit are always damped oscillations. Evidently, the damping rate will rise with the increase in  $R$ .

To make the oscillations in an actual circuit continuous there should be a device to compensate for energy losses similar to the mechanism replenishing the energy of the oscillating pendulum in a clock. In that case the excess energy of the oscillatory circuit,  $E$ , remains constant at the expense of some external power source and the oscillations will be continuous. A device capable of sustaining electrical oscillations is the *electron tube oscillator*.

Consider the principle of operation of the electron tube oscillator. Its schematic diagram is depicted in Fig. 30.2. The oscillatory  $L$ - $C$  circuit with a capacitor  $C$  and inductance  $L$  in which continuous oscillations have to be maintained is connected into the anode circuit of a triode, a coil  $L_1$ , inductively coupled with the coil  $L$ , is connected into the

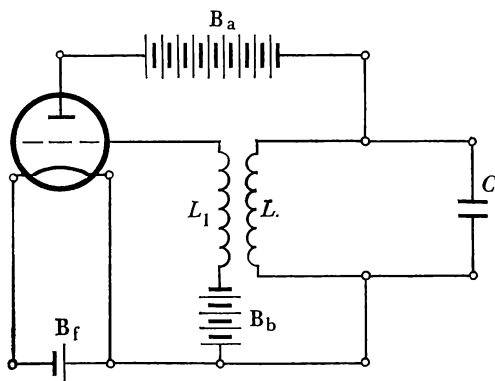


Fig. 30.2 Circuit diagram of electron tube oscillator:  $B_a$ , anode battery;  $B_f$ , filament battery;  $B_b$ , grid bias battery.

grid circuit. When the anode circuit is closed, the capacitor  $C$  is charged and electrical oscillations are excited in the  $L$ - $C$  circuit.

Because of the inductive coupling between coils  $L$  and  $L_1$  a voltage of the same frequency as in the  $L$ - $C$  circuit is induced in the grid circuit. The current in the anode circuit oscillates in time with the grid voltage. For the tube to replenish the energy of  $L$ - $C$  circuit the grid voltage should be in-phase with the anode current. In such conditions the tube automatically maintains the oscillations in the  $L$ - $C$  circuit at the expense of the energy of the anode battery.

This device can be used to produce high-frequency oscillations widely used in technology. The oscillation frequency in the  $L$ - $C$  circuit can be adjusted as required by a variable capacitor in place of  $C$  or by a variable inductance in place of  $L$ .

### 30-3 High-Frequency Currents

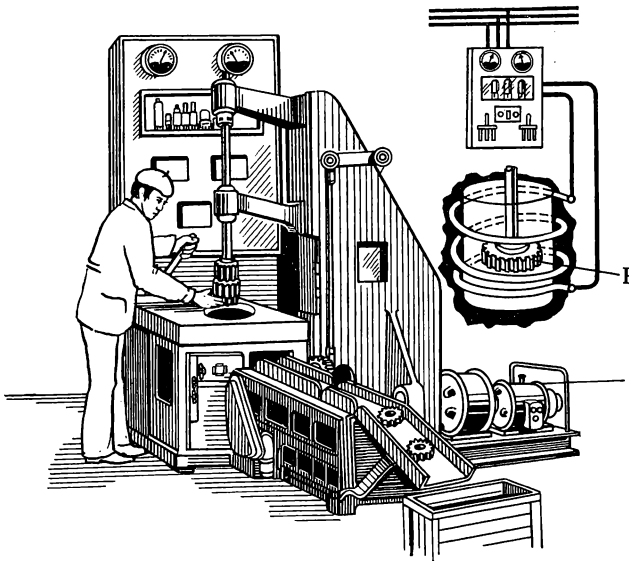
High-frequency currents have some peculiar features. When such a current flows in a conductor, eddy currents, due to rapid variations of the magnetic field, are excited inside the conductor.

The variations of the magnetic field inside the conductor are such that the eddy current near the conductor's axis is directed against the original current and at the conductor's surface coincides in direction with the latter. Hence, the distribution of high-frequency current over the conductor's cross section is nonuniform. At the centre of the conductor the current density is almost zero, then increasing from the centre to the external surface of the conductor.

A very high-frequency current flows practically only in a thin surface layer of the conductor. The term for this phenomenon is *skin effect*. For such currents thin tubes can be substituted for solid wires.

High-frequency currents are widely used at present. Let us cite several examples. There are furnaces used for

**Fig. 30.3** Surface hardening of metal products by high-frequency currents.



rapid heating and for melting metals. For instance, in the production of metal alloys containing volatile substances the metals are melted in special hermetic crucibles placed inside a coil fed with a high-frequency current.

The same procedure is used in the heat treatment of steel parts (Fig. 30.3). The part P is placed for a short time inside a coil fed with a high-frequency current. The part's surface layer is heated by eddy currents, the metal inside remaining cold. When the part is withdrawn from the coil, the cold metal inside quickly draws heat from the hot surface layer with the result that the latter cools down rapidly and

hardens. The depth to which the metal is heated in the process can be controlled by varying the time of heating and the current's frequency. The result of such heat-treatment is that the surface of the part is hardened, while the metal inside remains ductile.

To heat dielectrics they are placed inside a capacitor connected into an oscillatory circuit. The rapidly changing electric field causes vibrations of the dielectric's dipoles. This method is also used for drying wood and food products; in medicine it is used to heat ailing organs of the human body (diathermy).

### 30-4 Electromagnetic Field as a Special Form of Matter

It was stated in Section 26-7 that an alternating magnetic field creates a solenoidal electric field (see Fig. 26.8). The lines of this field are closed; it does not depend for its existence on electric charges and exists only as long as variations of a magnetic field continue. Its effect on electric charges is similar to that of an electrostatic field, the phenomenon of electromagnetic induction being proof of the fact.

Investigation of the relationship between electric and magnetic fields led Maxwell to base his theory of the electromagnetic field on two postulates:

- (1) an alternating magnetic field sets up a solenoidal electric field in the surrounding space;
- (2) an alternating electric field sets up a solenoidal magnetic field in the surrounding space.

Between the plates of a capacitor connected into an ac circuit there is an alternating electric field. This means that there should be a magnetic field as well. Therefore the variable electric field, from the point of view of its magnetic action, can be regarded as a sort of electric current without charges. To distinguish it from conduction current Maxwell introduced for it the term *displacement current*. Hence, applying the term electric current in its wider sense, we can say that the magnetic field is set up only by electric current and itself acts only on moving charges; the electric current, on the other hand, is set up by electric charges and by an alternating magnetic field and acts on all electric charges.

The variation of the electric field in a capacitor sets up a variable magnetic field in the surrounding space, which in turn sets up an electric field, and so forth. Hence, ev-

everywhere in space where variations of the fields take place there should be solenoidal electric and magnetic fields generating and sustaining each other. Because such fields are unseparable, it has been agreed to apply the term *electromagnetic field*.

It follows from this that if in some small region of space periodic variations of electric and magnetic fields are excited, such variations should be periodically repeated in other points of space as well. For a finite velocity of propagation of electromagnetic excitation there should be a time lag in these variations, increasing with the distance from the source. Thus, Maxwell's postulates lead to the conclusion that electromagnetic waves do exist in nature.

Applying his theory, Maxwell computed the velocity of propagation of electromagnetic waves in a vacuum and found it to be equal to the velocity of light  $c$  (see Section 31-6):

$$c \approx 3 \times 10^8 \text{ m/s} = 300\,000 \text{ km/s}$$

Since electric and magnetic fields both possess energy, there should be a specific amount of electric and magnetic energy in the space in which electromagnetic waves are propagating, this energy being carried by the waves from point to point in the direction of their propagation.

Experiments and subsequent development of Maxwell's theory proved the postulates cited above to be true.

Electromagnetic phenomena obey specific laws characterizing a special form of motion of matter, the electromagnetic form, which differs from the mechanical form of motion. Consider now the method of producing electromagnetic waves with the aid of an oscillatory circuit.

### 30-5 Open Oscillatory Circuit

Electrical oscillations are always accompanied by electromagnetic waves, but in practice these waves are not always easy to detect.

The main process that takes place in an oscillatory circuit, such as is depicted in Fig. 30.1, is the exchange of energy between the capacitance and the inductance, the energy lost on generating electromagnetic waves in space not being large. An oscillatory circuit with negligible radiative losses is termed a *closed oscillatory circuit* (see Section 30-1). What should be done to increase the intensity of radiated electromagnetic waves?

The first experiments in this field were carried out by Hertz (see Section 30-7), but the final solution to the problem was found only after the works of the Russian physicist Alexander S. Popov (1859-1905) and the Italian electrical engineer Guglielmo Marconi (1874-1937).

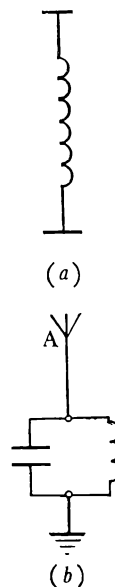
A closed oscillatory circuit does not set up electromagnetic waves in space, because the variations of the electric and magnetic fields are confined to a very limited space containing the inductance and the capacitance. To set up intensive waves oscillations should be induced in an open space with most of the variable fields enveloping the oscillatory circuit from all sides.

Note that the term for electromagnetic waves generated by an oscillatory circuit is *electromagnetic radiation*. To increase radiation one method is to draw the plates of the capacitor apart (Fig. 30.4a). But this method is not very effective.

Popov found a much more effective method of increasing the power radiated by an oscillatory circuit. He retained the original circuit, but grounded one end of the coil and connected a vertical wire with a free end to the second. The present term for this vertical wire A (Fig. 30.4b) is *feeder*. The device connected to an oscillatory circuit to increase the power of electromagnetic radiation and to receive electromagnetic waves (see Section 30-8) is called an *antenna*, or *aerial* (invented by Popov in 1895). An oscillatory circuit connected to an aerial is termed *open*.

Thus, to radiate electromagnetic waves an open oscillatory circuit is required.

Fig. 30.4 (a) Open oscillatory circuit; (b) oscillatory circuit with aerial.



### 30-6 Electromagnetic Waves

Periodic variations of electric and magnetic fields take place at every point of space where an electromagnetic wave is propagating. Such variations can conveniently be represented by the oscillations of the vectors  $\mathbf{H}$  and  $\mathbf{E}$  at every point in space.

Maxwell demonstrated that at every point the variations of these vectors are in-phase and take place along mutually perpendicular (orthogonal) directions (Fig. 30.5), which in turn are perpendicular to the wave propagation velocity vector  $\mathbf{v}$ . By way of an example, the relative arrangement of the vectors in the wave propagating from the aerial A is shown at point B. The relative arrangement of these vectors at any point of a travelling electromagnetic wave obeys the right-hand screw rule: if the screw head lying in the

Fig. 30.5 Electromagnetic wave propagates from point  $B$  in direction of  $\mathbf{v}$ .

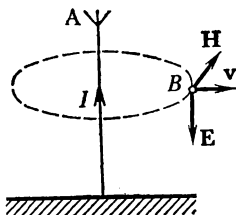
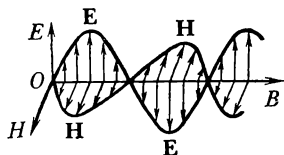


Fig. 30.6 Instantaneous orientation of vectors  $\mathbf{E}$  and  $\mathbf{H}$  in electromagnetic wave; wave propagates in direction  $OB$ .



plane of vectors  $\mathbf{E}$  and  $\mathbf{H}$  is turned in the direction of the shortest path from  $\mathbf{E}$  to  $\mathbf{H}$ , the translational motion of the screw will be in the direction of  $\mathbf{v}$ , that is, in the direction of the wave itself and of the energy transported by it.

Vectors  $\mathbf{E}$  and  $\mathbf{H}$  oscillate in a plane normal to vector  $\mathbf{v}$ . This means that the electromagnetic waves are transverse waves. The arrangement of vectors  $\mathbf{E}$  and  $\mathbf{H}$  at different points of the wave for the same instant of time is shown in Fig. 30.6.

The velocity of propagation of electromagnetic waves depends on the electric and magnetic properties of the medium, its numerical value, following Maxwell's theory, being

$$v = \frac{1}{\sqrt{\mu_m \epsilon_m}} \quad (30.6)$$

Since  $\mu_m = \mu \mu_0$  and  $\epsilon_m = \epsilon \epsilon_0$ , we have

$$v = \frac{1}{\sqrt{\mu \epsilon} \sqrt{\mu_0 \epsilon_0}} \quad (30.7)$$

Since the values of  $\mu$  and  $\epsilon$  in a vacuum are unit, the velocity of electromagnetic waves in a vacuum is

$$c = \frac{1}{\sqrt{\mu_0 \epsilon_0}} \quad (30.8)$$

(Demonstrate that (30.8) yields a value of  $c$  close to  $3 \times 10^8$  m/s.)

Comparing formulae (30.8) and (30.7) we obtain

$$v = \frac{c}{\sqrt{\mu \epsilon}}, \quad \text{or} \quad \frac{c}{v} = \sqrt{\mu \epsilon} \quad (30.9)$$

The term for the quantity  $n$  showing the number of times the velocity of propagation of electromagnetic waves in a vacuum exceeds that in a medium is the *absolute refractive index* of that medium:

$$n = c/v \quad (30.10)$$

The phenomena of wave refraction and the origin of the term for  $n$  are explained in Sections 32-4 and 32-5. Hence

$$n = \sqrt{\mu \epsilon} \quad (30.11)$$

Note that both the relation permittivity  $\epsilon$  and the relative permeability  $\mu$  in formula (30.11) are frequency dependent, decreasing with the frequency. For this reason, when making computations with the aid of formulae (30.6), (30.7), (30.9), and (30.11) one should avoid using data obtained from the tables of electrostatic values of  $\epsilon$ . The values of  $\mu$  for

all dielectrics with the notable exception of ferrites are close to unity. The value of  $\epsilon$  can never be less than unity. Therefore in all media the propagation velocity of electromagnetic waves is less than in a vacuum, or  $n$  always exceeds unity.

Formula (27.18) is valid for electromagnetic waves:  $v = \lambda f$ . For a vacuum this formula assumes the form

$$c = \lambda_0 f \quad (30.12)$$

where  $\lambda_0$  is the wavelength in a vacuum.

We recall that when a wave passes over from one medium to another, its frequency remains unchanged, but the wavelength changes. It should be kept in mind that wavelengths are always specified as for a vacuum, if not otherwise stated. In practice it is usual to use high-frequency waves because the energy radiated by a vibrator per unit time is proportional to the fourth power of the frequency. It is also easier to achieve directional radiation of electromagnetic waves at higher frequencies.

In communication, electromagnetic waves are often transmitted over wires (telegraphy), which serve, in a sense, as guides for the waves. Electric energy travels along wires with a velocity of  $3 \times 10^8$  m/s, that is, when the circuit is closed, the current is established practically simultaneously along the whole circuit, although the average speed of directional motion of the electrons in the wire is only a few millimetres per second.

### 30-7 Electrical Resonance

If an oscillatory circuit is placed in the path of electromagnetic waves, forced electromagnetic oscillations will be excited in it.

When the difference between the frequency of the electromagnetic waves and the natural frequency of a low-resistance oscillatory circuit is great, the amplitude of the forced oscillations in it is so small that it can be neglected. Noticeable electromagnetic oscillations are excited in the circuit only when the frequency of the forced oscillations in it coincides with its natural frequency, that is, when electrical resonance sets in. Hence, if one places several oscillatory circuits with different natural frequencies in the path of electromagnetic waves, intense electromagnetic oscillations appear only in the circuit "tuned in" with the radiator of those waves. The term for the excitation of oscillations in



Fig. 30.7 When two circuits with  $C_1$  and  $C_2$  are tuned in resonance, lamp L in circuit with  $C_2$  burns.

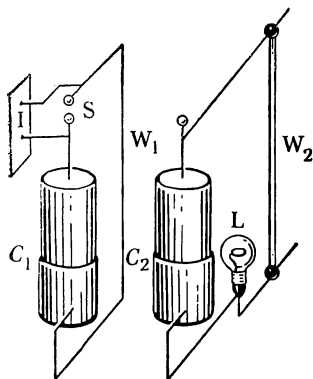
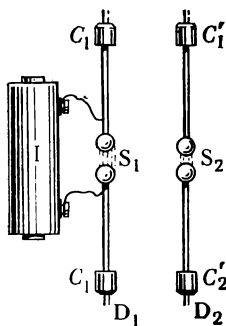


Fig. 30.8 Hertz's experiment; dipole  $D_1$  acts as vibrator and dipole  $D_2$  as resonator.



a resonator by the energy transmitted from an oscillator with the same frequency as the natural frequency of the resonator is *electrical resonance*.

Electrical resonance can be observed in an experiment that was first carried out by the British physicist Sir Oliver J. Lodge (1851-1940). In this experiment the capacitors (Fig. 30.7) were Leyden jars—glass jars covered within and without with lead foil to half the height of the jars. The internal foil of such a jar is connected to a rod bearing a ball. The oscillatory circuit in Lodge's experiment consists of a capacitor  $C_1$ , a wire  $W_1$ , and the spark-gap S, to which an induction coil I is connected in parallel (see Section 29-6). The resonator is a circuit made up of a capacitor  $C_2$ , a neon lamp L and a mobile wire  $W_2$ .

The circuits are placed parallel to each other and the induction coil I is switched on, causing a spark discharge in S in the process of which electromagnetic waves are radiated. By moving the wire  $W_2$  the second circuit is tuned in resonance with the first, the neon lamp L being used to control the tuning: when in resonance the lamp lights up. If we rotate the second circuit, we observe that when the circuits are at right angles to one another the lamp no longer burns. This makes it possible not only to detect electromagnetic waves but also to find the direction in which they travel.

Hertz was first to discover electromagnetic waves in a resonance experiment. The oscillatory circuits used by him were the so-called *dipoles* (Fig. 30.8). In Hertz's experiments, as in the experiments of Lodge, high-frequency power was supplied to the vibrator by an induction coil I. Hertz's dipole  $D_1$  was made of two wires,  $W_1$  and  $W_2$  with balls at their ends forming a spark gap. The opposite ends had caps  $C_1$  and  $C_2$  on them. By moving the caps one could change the capacitance of the circuit. The second dipole,  $D_2$ , was of identical design and was arranged parallel to the first. The spark in  $S_1$  caused dipole  $D_1$  to radiate electromagnetic waves, which were received by the resonator  $D_2$ . The resonator was tuned in resonance by moving the caps  $C_1'$  and  $C_2'$ . In resonance a spark appeared in the gap  $S_2$ , which could be detected with the aid of a strong magnifying glass.

The experiments described in this section prove that electromagnetic radiation can be transmitted and received. However, such methods of transmission and reception of radiowaves are not used at present, because such transmitters and receivers are very ineffective (a wide and unstable frequency spectrum of the transmitter and poor sensitivity of the receiver).

### 30-8 The Invention of Radio. Radiotelegraphy

Experiments by Hertz demonstrated the possibility of the transmission and reception of signals, but all this was done over very short distances, in fact within the limits of the laboratory table. Using an aerial Popov multiplied the power radiated by the vibrator and the sensitivity of the resonator and established long-distance communication with the aid of electromagnetic waves.

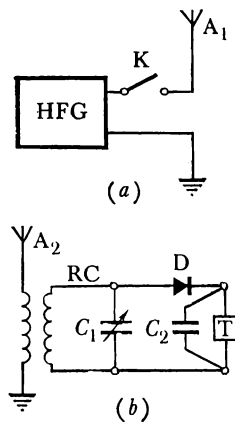
Having introduced improvements in the design of the transmitter and receiver, he started transmitting and receiving words using the Morse code. Very quickly he discovered that such signals could be received over earphones. The method became known as radiotelegraphy. Initially Popov was only able to establish communications over a few dozen metres, but in time he was able to transmit messages over a distance of several dozen kilometres. Popov's discoveries are very important. We all know the part played by radiocommunications in modern life.

The schematic diagram of a radiotelegraph communication system of the simplest type is depicted in Fig. 30.9. The transmitter is a generator of continuous high-frequency oscillations (HFG) connected via a key  $K$  to the aerial  $A_1$ . With the key closed the transmitter radiates electromagnetic waves. The receiver, with the aerial  $A_2$  connected to the resonance circuit (RC) and with a variable capacitor  $C_1$  used to tune the receiver in resonance with the transmitter, is far away from the latter. It is well known that in practice many transmitters are in operation at any one time. To prevent them from interfering with each other, every transmitter has to operate on its own frequency, different from the frequencies of the other transmitters. The capacitor  $C_1$  tunes the receiver in resonance with the specific transmitter.

The oscillations of the resonance circuit are transmitted via the detector  $D$  to the telegraph receiver (or to earphones)  $T$  or to a recorder. The detector (rectifier) transforms the high-frequency alternating current into a current flowing in one direction, that is, rectifies the high-frequency current. (We recall that rectifiers may be either of the vacuum-tube type or of the semiconductor type.) To smooth out pulsations of the rectified current in the telegraph receiver, a capacitor  $C_2$  is connected in parallel to it. It is charged during the time a current pulse passes through the receiver and partially discharged in the intervals between the pulses.

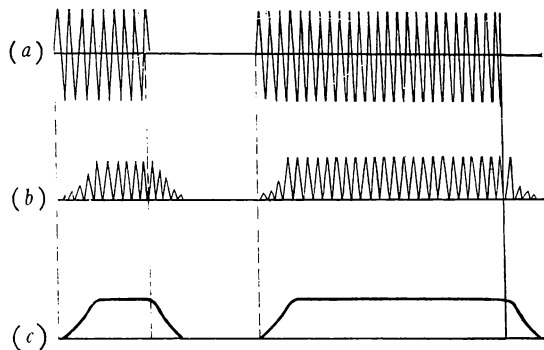
The transmission method is as follows. If we have to transmit a dot and a dash, first the key is pressed for a short time and then for a longer time. Two wave pulses

Fig. 30.9 (a) Block diagram of a radiotelegraph transmitter; (b) circuit diagram of simple receiver.

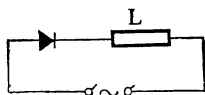


propagate from the transmitter: the short and the long (they are illustrated in Fig. 30.10a). Having passed through detector, the current pulses assume the shape shown in Fig. 30.10b and reach the telegraph receiver. The current pulses are smoothed out by the capacitor  $C_2$ , so that the

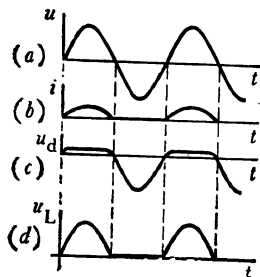
**Fig. 30.10** (a) Electrical oscillations in radio transmitter; (b) same oscillations after demodulation in receiver; (c) signals in telegraph recorder.



**Fig. 30.11** Connection of rectifier in series with load.



**Fig. 30.12** (a) Time dependence of voltage supplied by power source to terminals of circuit in Fig. 30.11; (b) current in circuit; (c) voltage variations across diode; (d) voltage variations across the load resistor.



current flowing in the telegraph receiver is of the shape shown in Fig. 30.10c. The term for the curve depicted in the figure is *envelope*, because it can be obtained by drawing a line tangent to all the pulse peaks in Fig. 30.10b.

The high-frequency signal must be rectified because the recording device of the telegraph receiver cannot operate at the rate corresponding to those of high-frequency oscillations. It is interesting to observe voltage variations in a circuit containing a rectifier. Figure 30.11 shows the connection of a semiconductor rectifier in series with a load  $L$ . The curve of the time dependence of the ac voltage in the main, or across the circuit's terminals is shown in Fig. 30.12a. The diode conducts current essentially only in the forward direction (Fig. 30.12b). We recall that in the case of a connection in series the voltage is distributed in proportion to the resistances. During the first half of the period, when the current in the diode flows in the forward direction, its resistance is quite small and almost the entire voltage drops on load  $L$ . During the second half of the period the diode's resistance is very great and the entire voltage is applied to the diode. The variations of the voltage across the diode are depicted in Fig. 30.12c, and across load  $L$  in Fig. 30.12d.

Telegraph signals can be received orally with the help of earphones. To this end the high-frequency signals have to be transformed into audio signals.

### 30-9 Amplitude Modulation. Radiotelephony

Radiocommunications in sound became possible only after vacuum amplifier tubes had been invented.

The difficulty in establishing sound communications lies in the fact that audio signals are low-frequency signals, and it is impossible to design efficient aerials for their transmission. This difficulty can be overcome by using radio-frequency (r.f.) waves as carriers of audio signals.

The variation of parameters (amplitude, frequency, phase) of r.f. oscillations over time with signals of a lower frequency is termed *modulation* of r.f. oscillations. The term for the r.f. wave being modulated is *carrier wave*, because it is used solely as an auxiliary device to transmit information contained in the signal of lower frequency (audio, television, etc.). The carrier frequency should be kept constant, that is, it should be well stabilized.

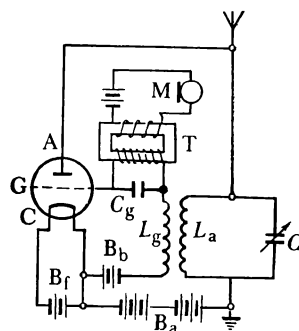
*Amplitude modulation* involves the variation of the carrier's amplitude with the frequency and the amplitude of the modulation signal. The following method can be used for amplitude modulation. A source of audio frequency oscillations is connected into the grid circuit of an electron tube oscillator producing continuous r.f. oscillations. Sound waves acting on the microphone M (Fig. 30.13) excite in its circuit audio frequency (a.f.) oscillations which are transmitted to the grid circuit of the vacuum tube via the transformer T.

Since the secondary of the audio transformer presents a great reactance to the r.f. signal it is bypassed by the capacitor  $C_g$ , with a low r.f. reactance. At the same time it does not short-circuit the audio signal because its audio-frequency reactance is high. The grid circuit also contains a grid bias battery  $B_b$ , which ensures that the grid potential remains negative with respect to the cathode even at high positive amplitudes of the r.f. signal.

In the absence of a.f. oscillations the transmitter operates as a generator of continuous r.f. oscillations (see Section 30.2) of constant amplitude. The variations of potential due to the microphone are led to the grid (Fig. 30.14a) and affect the anode-filament current and the amplitude of the high-frequency oscillations. Instead of oscillations with uniform amplitude, there is a controlled slow change of amplitude (Fig. 30.14b). This variation is modulation.

The modulated r.f. signal is received by the aerial of the receiver, is amplified and then rectified (Fig. 30.14c). The rectified a.f. signal is amplified and fed to earphones

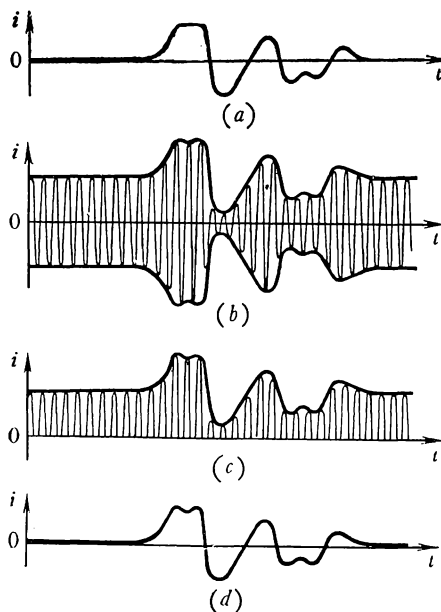
Fig. 30.13 Simple transmitter with modulator.



or loudspeaker (Fig. 30.14*d*), which transform the electric signal into sound.

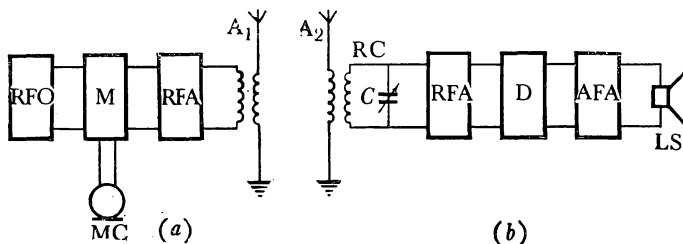
The schematic diagram of radiocommunications system presented in Fig. 30.15 shows the main units making up

**Fig. 30.14** (a) Audio-frequency oscillations; (b) amplitude modulated oscillations; (c) demodulated oscillations; (d) audio-frequency oscillations in earphones.



a transmitter and a receiver. The first unit of a transmitter is the r.f. oscillator, RFO, the second is the modulator, M, where the signals from the microphone, MC, modulate

**Fig. 30.15** Block diagram of (a) transmitter and (b) radio receiver.



the r.f. carrier wave, the third is the r.f. amplifier, RFA, and the fourth the transmitter aerial,  $A_1$ .

The first unit of the receiver is the aerial  $A_2$ , the second the resonance circuit RC, the third the r.f. amplifier, RFA, the fourth the detector, D, which rectifies the modulated

r.f. signal, the fifth the a.f. amplifier, AFA, and finally the loudspeaker, LS.

Modern vacuum tube and transistor amplifiers can amplify r.f. signals by a factor of several millions. Thus modern radio receivers can receive transmissions from radiostations all over the world.

### 30-10 A Simple Vacuum Tube Receiver

The circuit diagram of a vacuum tube receiver of the simplest type is depicted in Fig. 30.16. Let us see how it works.

The radio waves from the distant transmitter impinge on the receiving aerial and cause forced oscillations in it. The inductive coupling between the coils  $L_1$  and  $L_2$  transmits these oscillations to the oscillatory circuit  $L_2$ - $C_1$

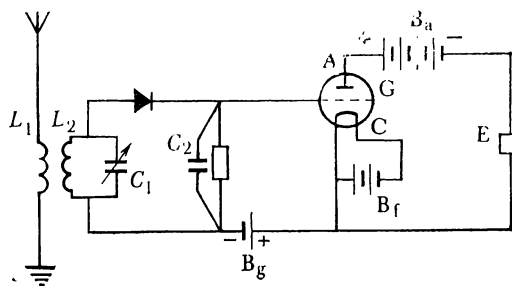
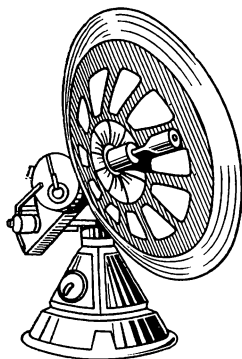


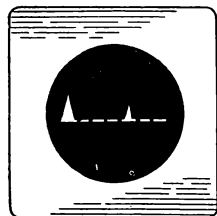
Fig. 30.16 Circuit diagram of simplest vacuum tube radio receiver.

tuned in resonance by means of the capacitor  $C_1$ . The oscillations in this circuit are, after detection, applied to the grid of the triode, causing substantial variations in its anode-filament current, along with the variations in the grid voltage. The amplified a.f. signal acts on the earphone membrane, E, which transforms electric oscillations into mechanical vibrations, that is, produces sound waves. If the sound produced in the earphones is not loud enough, additional amplification can be introduced. To this end a load resistor is inserted into the anode circuit of the first triode instead of the earphones, and the signal from the anode is transmitted (via a dividing capacitor) to the grid circuit of the second triode, the earphones being connected into the anode circuit of the latter. All modern radio receivers have an r.f. amplifier preceding the detector (see Fig. 30.15) to enable them to receive weak signals from distant stations.

**Fig. 30.17** Directional antenna of radar system.



**Fig. 30.18** Signals on radar screen.



### 30-11 Radar

Section 28-8 dealt with the ultrasonic direction and range finder (sonar). There is a similar device operating with electromagnetic waves.

Popov, while conducting his radiocommunication experiments, discovered in 1897 the scattering of radiowaves by a ship. This phenomenon forms the basis of *radar* (short for radio direction and ranging).

The basic principle of radar is the reception of the electromagnetic “echo” and, accordingly, such a set must radiate electromagnetic waves and receive them back after they are scattered by the object. Its radiation must be strictly directional. The set must include a device for the accurate measurement of intervals between the time a pulse is sent and the time it is received after being scattered.

Since it is easier to realize directional radiation with short waves, very high-frequency waves (vhf), for example centimetre waves, are used for radar. A radar station is equipped with a highly directional antenna system, its shape resembling that of a reflector, with a vhf emitter mounted in the centre (Fig. 30.17). The same antenna serves for the reception of pulses scattered by the obstacle.

A radar station has a cathode-ray tube which receives signals when the pulses are sent and received. The resulting pattern on the screen is shown in Fig. 30.18. The interval between the time the pulse was sent and the time it was received can be found from the known time the electron beam crosses the screen’s diameter. The distance to the obstacle is found by multiplying the velocity of wave propagation,  $3 \times 10^8$  m/s, by half the time interval between the transmission and the reception of the pulses (why multiply by a half?). Usually the scale on the screen is graduated directly in kilometres (or miles) and the transmission moment is displaced to the origin of the scale so that the distance to the obstacle is determined by the position of the received pulse.

Radar is widely used in practice: in aircraft to measure altitude and for landing in conditions of poor visibility, on ships to detect obstacles, in astronomy to measure distances to celestial bodies, etc.

### 30-12 The Cathode-Ray Oscilloscope

The *electronic*, or *cathode-ray*, *oscilloscope* is designed for the study of high-speed processes. The operating panel of a typical oscilloscope is shown in Fig. 30.19.

It operates on the following principle. One of its principal parts is the cathode-ray tube (see Section 23-11). In a tube with electrostatic deflection a voltage changing linearly with time from zero (this voltage corresponds to the beam on the extreme left of the horizontal diameter of the screen) to the maximum (corresponding to the extreme

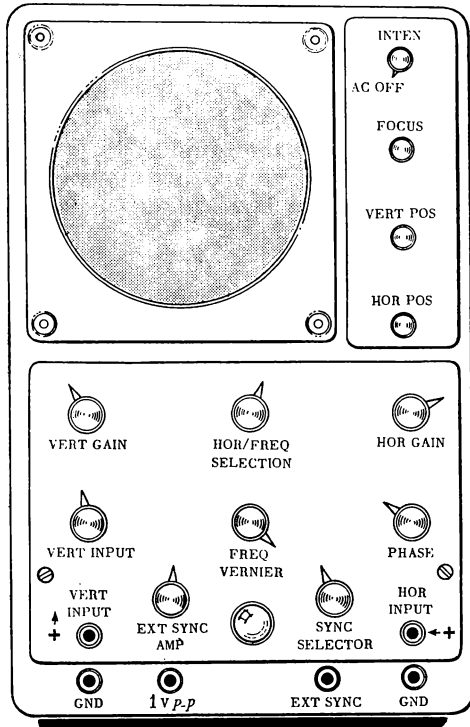


Fig. 30.19 Operating panel of cathode-ray oscilloscope.

right position of the beam) and then dropping instantly to zero is applied to the horizontal deflection plates. When the deflection voltage reaches the maximum value, the beam is interrupted, the deflection voltage jumps to zero and the beam returns to the extreme left position. The process is repeated at equal intervals, a device called a *variable time base* being used to vary the magnitude of the intervals.

Suppose the time base is set at 0.02 s. If we now apply to the second pair of deflection plates, which deflect the beam in the vertical direction, a voltage from an ac main, a sinusoid will be displayed on the screen, because the period of the ac main voltage is also 0.02 s.



In the absence of a signal on the vertical deflection plates a horizontal straight line will be displayed on the screen. The voltage which deflects the beam in the horizontal direction is termed *sweep voltage*, the name for the device which produces such voltage being *sweep oscillator*. The plot displayed on the oscilloscope's screen is called an *oscillogram*.

The oscilloscope can also be used to study high-speed mechanical processes, for instance mechanical vibrations. To this end mechanoelectric transducers, which transform the energy of mechanical vibrations into electric signals, for instance piezocrystals, are used. The output of such a transducer is applied to the input of the vertical deflection amplifier of the oscilloscope and in this way an oscillogram of mechanical vibrations is obtained.

part four

# **Optics and Special Relativity**

# The Nature of Light.

## Propagation of Light

### 31-1 Historical Survey

The part of physics dealing with light phenomena is termed *optics* (from the Greek *optikos* for visual) and the phenomena themselves are termed *optical*.

Light falling on objects makes it possible for us to see them and to find our way in space.

This, however, is not the only effect of light. Recall, for instance, how hot objects become in the sun. This means that light has energy and transports it through space. Since energy can be transported either by bodies or by waves, there are two possible hypotheses about the nature of light. Luminous radiation must be either a flux of tiny particles, which Newton termed *corpuscles*, or a train of waves spreading in some medium.

Newton used the first hypothesis to create his *corpuscular theory of light*, which was used to explain numerous optical phenomena. For instance, the difference in colour was explained as the difference in the shape of its corpuscles. The second hypothesis served as the basis for the *wave theory of light* proposed by the Dutch scientist Christian Huygens (1629-1695) in the second half of the seventeenth century. Huygens' theory successfully explained such phenomena as the interference and diffraction of light.

Since neither theory was alone able to explain all optical phenomena, the question as to the true nature of light remained unanswered. At the beginning of the nineteenth century the results obtained by Fresnel, Foucault and many other scientists put the wave theory ahead of the corpuscular. However, there was one point that remained unclear. The wave theory presumed light to be a form of transverse (this was proved by experiment) mechanical waves. Therefore there must be a substance in the space between the Earth and the Sun, since light easily reaches the Earth from the Sun. This led to the hypothesis of a *universal ether* which fills all space between bodies and molecules. If one recalls that transverse mechanical waves are only possible in solids (see Section 27-15), one has to concede that such ether must possess the properties of an elastic solid and yet in no way affect the motion of the Earth through space. This means that the ether makes itself manifest only in transporting light and gravitation, although it has the properties of a solid. Such controversial properties of the ether cast doubt on its existence.

This inconsistency of the wave theory of light was, on the whole, eliminated by Maxwell. Having drawn up his electromagnetic theory, Maxwell noted that the theoretical value of the velocity of propagation of electromagnetic waves in a vacuum coincides with the velocity of light measured experimentally. This prompted him to advance a hypothesis on the electromagnetic nature of light, which was subsequently substantiated by numerous experiments.

Thus, towards the end of the nineteenth century the *electromagnetic theory* of light came into being and is still in use.

### 31-2 The Electromagnetic Theory of Light

According to the electromagnetic theory of light any luminous radiation is electromagnetic waves. However, by no means all electromagnetic waves are light waves, but only such as can be seen by a human eye. Luminous radiation includes only waves with frequencies lying in the range from  $4 \times 10^{14}$  to  $7.5 \times 10^{14}$  Hz. There is a separate colour to correspond to every frequency band inside this interval. For instance, the colour corresponding to the frequency of  $5.4 \times 10^{14}$  Hz is green. From the frequency of the radiation one can find its wavelength in a vacuum, using formula (30.12):

$$\lambda = c/f$$

Accordingly the wavelengths of luminous radiation in a vacuum range from 400 nm\* (violet) to 760 nm (red). Note that when luminous radiation enters another medium, its colour remains unchanged since its frequency remains constant, but the wavelength changes because of a change in the propagation velocity of light. When wavelength is used to characterize the colour of radiation, the wavelengths indicated are for a vacuum.

Maxwell concluded on the basis of his theory that luminous radiation, as well as other electromagnetic waves, must exert pressure on bodies. P.N. Lebedev proved this by his experiments (see Section 38-2).

### 31-3 The Quantum Theory of Light

It was established as a result of the analysis of the spectral (frequency) distribution of radiation of luminous bodies that existing theories, thermodynamics and the electromagnetic theory of light, were unable to explain this distribution. In an effort to find an explanation the German physicist Max Planck (1858-1947) suggested that light is radiated in definite indivisible portions of energy which he termed *quanta* (from the Latin *quantus* for how great). The present term for light quanta is *photons*.

The analysis of optical phenomena led to the conclusion that the phenomena involving the propagation of light in a medium could be adequately explained on the basis of wave theory, but the phenomena involving the emission and absorption of light could not be explained without the concept of the quantum nature of luminous radiation. This meant that a new theory which combined the wave and the corpuscular properties of light was required. This new theory became known as the *quantum theory of light* and in its original form was the result of the work of Planck, Einstein, Bohr and other scientists.

At present quantum theory explains not only optical phenomena but also numerous other phenomena from all fields of physics. This theory discovered new properties of substances and fields and predicted many new phenomena which were subsequently discovered in experiment.

The relation between the wave and corpuscular properties of light is expressed by Planck's formula:

$$\varepsilon = hf \quad (31.1)$$

\* 1 nm (nanometer) is equal to  $10^{-9}$  m.

where  $\varepsilon$  is the quantum's energy,  $f$  is the frequency of oscillations of the electromagnetic radiation and  $h$  is a constant identical for all waves and quanta and termed *Planck's constant*. The value of  $h$  in the SI system is

$$h = 6.62 \times 10^{-34} \text{ J}\cdot\text{s}$$

According to quantum theory, a luminous radiation of specified frequency  $f$  is made up of photons (quanta) of definite energy  $\varepsilon$  expressed by formula (31.1). Hence, the energy of a quantum is directly proportional to the frequency of the electromagnetic radiation. Since  $c = f\lambda$ , we obtain from formula (31.1)

$$\varepsilon = hc/\lambda \quad (31.2)$$

that is, the energy of a quantum is inversely proportional to the radiation's wavelength in a vacuum.

It was established experimentally that as long as a photon exists it can move in a vacuum only with a velocity equal to  $c$  and in no circumstances can it be either accelerated or decelerated or stopped. On meeting a substance it can be absorbed by a particle of the substance. In that case the photon itself disappears, its entire energy being transmitted to the particle which absorbed it. The photon has no rest mass. This remarkable peculiarity of the photon distinguishes it from particles of substance such as the proton or electron.

Note that it is still not quite clear, from the classical viewpoint, why light in some phenomena displays pronounced wave properties while displaying corpuscular properties in other and how such contradictory properties can be combined in radiation. According to quantum theory such a combination of corpuscular and wave properties is a natural quality of matter as a whole, that is, every particle has wave properties and every wave has corpuscular properties.

### 31-4 Sources of Light

All bodies whose molecules and atoms emit visible radiation are termed light sources. Numerous examples of light sources can be cited: incandescent lamps, burning matches, gas-discharge tubes, etc. They can be classified by the method their light-radiating particles are excited.

The first group includes *thermal light sources*, whose radiation is the result of the excitation of atoms and molecules by random motion of the particles in the body at

appropriate temperatures. The radiation energy of such light sources is obtained at the expense of their internal energy.

The second group includes *luminescent light sources* whose atoms and molecules are excited not by high temperatures but by a flux of energetic particles, for instance by electrons, by external electromagnetic radiation or by chemical reactions. In this case the radiation energy is obtained at the expense of electrical, chemical or mechanical energy, that is, at the expense of some external source of energy. The screen of a cathode-ray oscilloscope, a gas-discharge tube in a luminous state, luminescent paints, etc. are examples of luminescent light sources. The phenomenon of luminescence will be discussed in more detail in Section 38-17.

The third group includes sources of *Cherenkov radiation*. This radiation appears when electrons (or other charged particles) move in a substance at speeds exceeding the speed of light in that substance (see Section 40-5).

### 31-5 Huygens' Principle

Let us see how wave theory explains the progress of a wavefront through space.

Let the front of a spherical wave propagating from point  $O$  at some instant of time be in position  $I$  (Fig. 31.1). After a period of time it will move to position  $II$ . The progress of the wavefront through space is explained with the aid of *Huygens' principle*: all points on the wavefront are oscillators from which elementary waves propagate ( $I$ ,  $2$ ,  $3$ , etc. in Fig. 31.1), while the envelope of all those elementary waves is the new wavefront (surface  $II$ ). (The envelope is a surface tangent to all the elementary waves.)

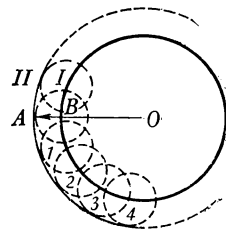
It should be taken into account at this point that the superposition of waves travelling in the direction of point  $O$  results in mutual suppression of oscillations, that is, the waves cancel each other in this direction.

The direction of motion of the wavefront in Fig. 31.1 is indicated by the arrow from  $B$  to  $A$ . We recall that the line along which the wavefront travels is called a *ray* (see Section 27-16). In a homogeneous isotropic medium, light propagates along straight lines. Many phenomena are proof of this, for instance the shape of shadows of non-transparent objects placed in the path of light rays. (Cite two more examples in support of the contention that light propagates along straight lines.)

As the distance of the wavefront from point  $O$  (Fig. 31.1) increases, its radius of curvature diminishes. Therefore at great distances from the light source a small area of a spherical wavefront may be regarded as a plane and the light rays may be assumed to be parallel. For instance, the sun's rays impinging on the Earth are assumed to be parallel.

For the sake of brevity we will speak of energy and colour of a ray, meaning the energy and colour of the radiation propagating in the direction of that ray.

Fig. 31.1 Propagation of wavefront according to Huygens' principle (arrow indicates direction of propagation of front along line  $OB$ ).



### 31.6 The Velocity of Light in a Vacuum

Because of a high velocity of propagation the travelling time of light is perceptible only over great distances; for instance, light travels from the Sun to the Earth in about 8 minutes.

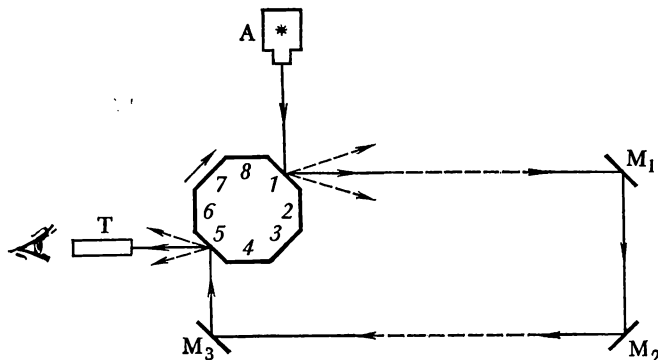
In 1675 the Danish astronomer Ole Roemer (1644-1710) was the first to measure the velocity of light in a vacuum. Observing the eclipse of one of Jupiter's satellites, he noticed that as the distance between Jupiter and the Earth increased, the satellite's eclipse was increasingly behind calculated time. Roemer's explanation was that as the distance between Jupiter and the Earth increased by  $l$ , light took time  $t$  to travel this distance with the velocity  $c$ . Knowing  $l$  and  $t$ , he computed the velocity of light, which proved to be close to  $3 \times 10^8$  m/s.

Subsequently the velocity of light was measured on numerous occasions in quite different conditions. A very accurate result for the velocity of light in air was obtained by the American physicist Albert A. Michelson (1852-1931). Let us consider one of his most successful experiments (made in 1926).

Michelson mounted an octagonal mirror (with  $k = 8$  faces) on a platform that could be rotated (Fig. 31.2). The light from an arc lamp  $A$  was directed at one of the faces. After being reflected by the face and by the mirrors  $M_1$ ,  $M_2$  and  $M_3$ , it fell on another face, and then, being reflected by it, reached the eye of the observer. The total distance  $l$  from the drum to the mirrors  $M_1$  and  $M_2$  and back to the drum, about 70 km, was accurately measured. The observer adjusted his telescope  $T$  so that he was able to see the image of the light source  $A$  clearly, and then set the platform in rotation. The image of the light source disappeared. (Why?) Gradually increasing the speed of rotation, the observer is able, at a speed of  $n$  r.p.m., to again see a clear image of  $A$ . This means that the drum turns exactly through the



Fig. 31.2 Michelson's octagonal-mirror experiment for measuring light velocity.



angle between the adjacent faces during the time the light travels between the mirrors. Since this time can be expressed by the formula

$$t = \frac{1}{kn/60} = \frac{60}{nk}$$

we obtain for the velocity of light  $c$  the formula

$$c = l/t = lnk/60$$

Since  $c$  is the maximum speed of signals in nature and is in numerous formulae, its value is one of the most important physical constants. After numerous checks it was established that

$$c = 299\,792.5 \pm 0.3 \text{ km/s}$$

### 31-7 The Velocity of Light in a Medium

It was stated in Section 30-6 that the velocity of propagation of electromagnetic waves depends on the medium and is expressed by the formula (30.6)

$$v = \frac{1}{\sqrt{\mu_m \epsilon_m}}$$

It was established that light can propagate only in dielectrics. Since the magnetic permeability for most dielectrics is very close to unity,  $\mu_m = \mu_0$ , the velocity of light in them is determined by their permittivity  $\epsilon$ . Because of the dependence of  $\epsilon$  on the frequency of oscillations of the vector  $\mathbf{E}$ , the velocity of light in dielectrics depends on the frequency of luminous radiation.

In 1850 Foucault was the first to measure the velocity of light in water. The quantity characterizing the dependence

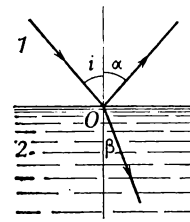
of the velocity of light on the medium is termed *optical density* of the medium. The measure for it is the value of the absolute refractive index of the medium  $n$  (see Section 30-6):

$$n = c/v$$

The optical density of a vacuum is obviously unity. Since the refractive index for air is  $n = 1.003$ , the velocity of light in air is often taken to be  $c$ .

The change in the propagation velocity of light is the cause of the refraction of light, that is, of the change in the direction of its propagation accompanying its passage from one medium to another.

Fig. 32.1 Reflection and refraction of light on boundary of two transparent media with different optical densities.



## Reflection and Refraction of Light

# 32

### 32-1 Optical Phenomena at the Boundary Surface Between Two Media

We recall that in a homogeneous and isotropic medium light propagates along a straight line. This allows us to use light rays to describe the propagation of light in such a medium.

At the boundary separating two media the direction of light propagation changes. Hence, if we find the laws governing such changes from experiment, we will be able to describe many optical phenomena without having to look into the physical nature of luminous radiation. The section of optics which uses such an approach to phenomena is termed *geometrical optics*. This chapter deals with laws governing optical phenomena at the boundary separating two transparent media.

When a narrow light beam falls from air onto the surface of water (Fig. 32.1), some of the light can be seen to be reflected at the point of incidence  $O$  and some to penetrate into water. This light undergoes refraction. We recall that the terms for the angles  $i$  and  $\alpha$  are angle of incidence and angle of reflection (see Section 27-19). The term for the angle  $\beta$  between the refracted ray and the normal to the boundary surface at the point of incidence is *angle of refraction*.

One could ask what part of the energy transported to the boundary surface separating two media will be carried

away by the reflected rays and what part by the refracted rays? Let the radiation energy transported to point  $O$  in a certain interval of time be  $E$ . Here this energy is divided: one part goes with the reflected rays ( $E_{rl}$ ) and the other with the refracted ( $E_{rr}$ ). It follows from the law of energy conservation that

$$E = E_{rl} + E_{rr}$$

Since any medium with the exception of a vacuum absorbs radiation energy, this equality holds only in the vicinity of point  $O$ . If luminous radiation is able to travel long distances in a medium without losing much of its intensity, the medium is said to be *transparent* (for instance, glass, water, alcohol, etc.). Metals, on the other hand, strongly absorb luminous radiation, that is, they are not transparent with regard to it. Metals reflect the greater part of incident radiation.

Thus, all media to some extent reflect and absorb luminous radiation. The quantity characterizing the reflectivity of a surface is called the *reflection factor*. The measure for the reflection factor is an abstract number equal to the ratio of energy carried away from a surface by reflected radiation to the energy transported to it:

$$r = \frac{E_{rl}}{E} \quad (32.1)$$

The reflection factor depends on the structure of the surface, the composition of the radiation, the angle of incidence and many other factors. However, if (other conditions remaining constant) only the angle of incidence is varied, an increase in angle  $i$  is accompanied by an increase in the reflection factor (see Table 32.1).

**Table 32.1** Percentage of energy reflected from glass and passing through it at the air-glass boundary for different angles of incidence

Angle of incidence (deg)	0	10	20	30	40	50	60	70	80	89
Reflected energy	4.7	4.7	4.7	4.9	5.3	6.6	9.8	18	39	91
Passing energy	95.3	95.3	95.3	95.1	94.7	93.4	90.2	82	61	9

Note that the frequency dependence of reflection and absorption is usually a selective one, that is, a body reflects radiation of some frequencies better than others; the same

is true of absorption. For instance, the atmosphere of the Earth strongly absorbs short waves of the visible spectrum and to a much lesser extent the longer waves of the spectrum. There are no surfaces that reflect all incident radiation. The surfaces of soot and of black velvet are practically nonreflecting, while a polished silver surface is an almost ideal reflector.

The term used for a body which absorbs all incident radiation is *black body*. It follows from the above that no black bodies really exist. However, it is possible to construct a perfect model of a black body (see Section 37-10). The black body is the most appropriate model for deducing theoretical laws of thermal radiation and of its absorption.

(Why is red light used for warning signals and not green light, to which the sensitivity of the human eye is greater?)

### 32-2 The Laws of Light Reflection

Experimental laws of light reflection were discovered as far back as the third century B.C. by the Greek scientist Euclid. In modern conditions these laws can be verified with the aid of an optical disk (Fig. 32.2) consisting of a light source *L* which can be moved along the disk's circumference, on which degree divisions are marked. The light is directed at the reflecting surface *M*, and the angles *i* and  $\alpha$  are measured.

The laws of light reflection are the same as those of the reflection of waves from obstacles (see Section 27-19):

(1) Incident and reflected rays lie in the same plane as the normal to the reflecting surface at the point of incidence.

(2) The angle of reflection is equal to the angle of incidence:

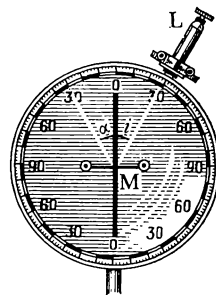
$$\angle \alpha = \angle i$$

With the aid of the optical disk one can prove that the incident and reflected rays are reversible, that is, if the incident ray is directed along the path of the original reflected ray, the new reflected ray will take the path of the original incident ray.

In Section 27-19 the laws of reflection were established for a spherical wavefront. Let us now demonstrate that they are also valid for a plane wavefront, that is, for the case of a parallel light beam falling on a plane surface.

Let a plane wave whose front at a definite moment of time coincides with the plane  $A_1B_1$  fall on a smooth plane surface  $KM$  (Fig. 32.3). The next moment the wavefront

Fig. 32.2 Optical disk.





takes place from all rough surfaces, for instance from the walls of a room. Light scattered by various nonluminous bodies makes them visible to us.

An ideally smooth nontransparent surface is termed a *mirror surface*. There are different types of mirrors besides the plane mirror: spherical, parabolic, etc. A parallel group of rays remains parallel after reflection from a plane mirror but changes the direction of its propagation (Fig. 32.5). Such reflection is termed *mirror*, or *regular*.

The definition of smooth and rough surfaces is not absolute but depends on the wavelength of the incident radiation. A surface is *smooth* if the details of its relief are smaller in linear dimensions than the wavelength of the radiation.

Light rays from a bright source reflected by a plane mirror may cause pain and temporary blindness on reaching the eye. In similar circumstances scattered light causes no discomfort. However, in other circumstances (for instance, in the mountains) the effect of light scattered by snow may be no less adverse and lead to snow blindness.

If light scattered by the surfaces of various bodies falls on a plane mirror and then after reflection reaches the human eye, man sees the images of the bodies in the mirror. Let us find out the origin of those images. We will begin with the image of a luminous point in a plane mirror.

Let there be a point light source  $S$  above the surface of the mirror  $KM$  (Fig. 32.6). The ray  $SA$  falling from the source at right angles to the mirror will, after reflection, reverse its direction, that is, it will take the path  $AS$ . From all the rays falling on the mirror from  $S$  choose the ray  $SB$  which falls on the mirror at the angle  $i$ . After reflection it takes the path  $BD$  so that  $\sphericalangle \alpha = \sphericalangle i$ . It is seen in Fig. 32.6 that the rays meeting the mirror at points  $A$  and  $B$ , take, after reflection, the paths they would have taken if they had propagated in the absence of the mirror from the same point  $S_1$  symmetrical about  $S$  with respect to the mirror  $KM$ . Let us prove this.

The angle  $\varphi$  is equal to the angle  $\alpha$ ; therefore  $\sphericalangle i = \sphericalangle \varphi$ . Since  $CB \perp KM$ , it follows that  $\sphericalangle 1 = 90^\circ - \sphericalangle i$  and  $\sphericalangle 2 = 90^\circ - \sphericalangle \varphi$ , or  $\sphericalangle 1 = \sphericalangle 2$ . This means that the right triangles  $SAB$  and  $S_1AB$  are equal because they have a common leg  $AB$  and equal acute angles  $1$  and  $2$ . Therefore  $SA = S_1A$ . This equality remains true for all rays falling from  $S$  on the mirror. (Prove this.)

Hence, when a man looks into a mirror, he sees the image of the light source  $S$  at point  $S_1$ , although there are actually no rays radiated from point  $S_1$  and reaching his eyes. Such an image is said to be *virtual*. If a screen is placed at point

Fig. 32.4 Scattered (diffuse) reflection.



Fig. 32.5 Regular (mirror) reflection.

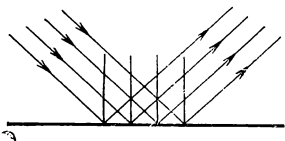
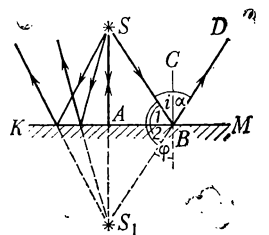
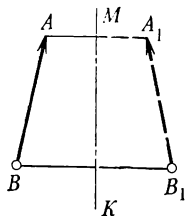


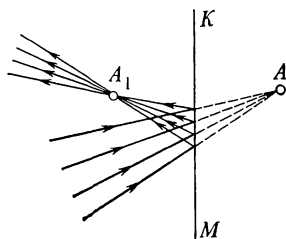
Fig. 32.6 Virtual image of real luminous point.



**Fig. 32.7** Mirror image of object.



**Fig. 32.8** Real image of virtual luminous point.



$S_1$  where the man sees a luminous point, there will be no image of point  $S$  on that screen. This is a property peculiar to the virtual image. In all other respects except its position the virtual image is quite authentic.

Thus, a plane mirror produces a virtual image of a luminous point  $S$  at point  $S_1$  symmetrical to it about the plane of the mirror.

Now imagine that there is an object in front of the mirror, schematically depicted in Fig. 32.7 as the arrow  $BA$ . The image of this object in the mirror can be found as follows. Draw normals from the extreme points of the object to the plane of the mirror and continue them beyond that plane to distances equal to those from the points to the plane; we obtain points  $A_1$  and  $B_1$ . Connect these points by a straight line and obtain the image of the arrow  $BA$  in the mirror. This image will be a virtual full-scale one. It has one peculiarity which distinguishes it from other images: the left-hand and right-hand sides of the image in the mirror are reversed as compared to the object. The term for such an image is *mirror image*.

There are not only virtual images but virtual light sources as well. Place a plane mirror  $KM$  (Fig. 32.8) in the path of rays which converge at point  $A$  (such rays can be obtained with the aid of a lens). In this case the reflected rays will converge at point  $A_1$  and then proceed as a diverging group, that is, an observer to the left of  $A_1$  will see point  $A_1$  as a real image of some point source of light. In contrast to a real point source from which the light rays diverge, the point to which their continuations converge is termed a *virtual light source*. In our case it is point  $A$ .

Thus, in a plane mirror the image of a real light source is virtual and located behind the mirror, and the image of a virtual light source is real and located in front of the mirror.

### 32-4 The Laws of Light Refraction

It was pointed out above that the refraction of light is due to the change in the velocity of its propagation taking place as light travels from one medium to another. Let us discuss in more detail the phenomena of light refraction in terms of wave theory.

Let a beam of parallel rays  $A'B'$  whose front is initially at  $AC$  fall on the boundary  $KM$  separating two transparent media (Fig. 32.9). Let the propagation velocity be  $v_1$  in the first medium and  $v_2$  in the second and let  $v_1 > v_2$ . Then

in the time  $t$  in which the wavefront travels the distance  $CB = v_1 t$  in the first medium the light propagating from point  $A$  in the second medium will reach the surface of the hemisphere with a radius  $AD = v_2 t$ . Accordingly, the wavefront by that time will occupy position  $BD$  and henceforth will travel parallel to itself in the direction  $AA''$ , or  $BB''$ .

Thus, if  $v_1 > v_2$ , the rays crossing from the first medium into the second are refracted so that the angle of refraction

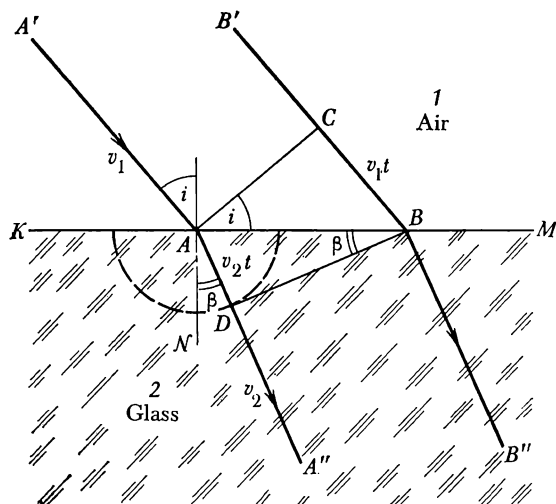


fig. 32.9 According to Huygens' principle, first elementary wave starts propagating in glass from point  $A$ , other elementary waves propagating in succession from points lying on line  $AB$  when elementary wave in glass starts propagating from point  $B$ , tangent to all elementary waves is in position  $DB$ .

is less than the angle of incidence, that is, the rays draw closer to the normal  $AN$ .

Let us find the relation between the angles  $i$  and  $\beta$ . From the right triangle  $ABC$  we obtain

$$BC = AB \sin i$$

and from the triangle  $ABD$

$$AD = AB \sin \beta$$

Dividing these equations term by term we obtain

$$\frac{BC}{AD} = \frac{\sin i}{\sin \beta}$$

Since

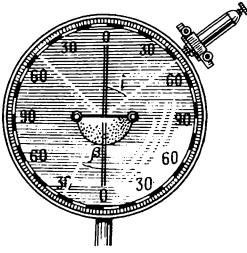
$$\frac{BC}{AD} = \frac{v_1 t}{v_2 t} = \frac{v_1}{v_2}$$

it follows that

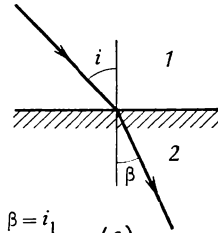
$$\frac{\sin i}{\sin \beta} = \frac{v_1}{v_2} \quad (32.2)$$



**Fig. 32.10** Checking light refraction laws with optical disk.

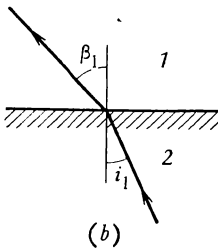


**Fig. 32.11** Reversibility of refracted light rays.



$$\beta = i_1$$

$$i = \beta_1$$



The ratio of light propagation velocities in both media is denoted  $n_{21}$  and termed the *refractive index of the second medium with respect to the first*, or *relative refractive index*:

$$n_{21} = \frac{v_1}{v_2} \quad (32.3)$$

Comparing formulae (32.3) and (32.2), we obtain

$$\frac{\sin i}{\sin \beta} = n_{21} \quad (32.4)$$

Formula (32.4) can be verified experimentally with the aid of the optical disk (Fig. 32.10). The experiment will also convince us that the incident and refracted rays lie in one plane with the reflected ray.

Hence, there are two laws governing light refraction:

(1) Incident and refracted rays lie in the same plane as the normal to the boundary separating the two media at the point of incidence.

(2) The ratio of the sine of the angle of incidence to the sine of the angle of refraction for two specified media is independent of the angle of incidence:

$$\frac{\sin i}{\sin \beta} = n_{21}$$

It follows from the second law that an increase in the angle of incidence results in an increase in the angle of refraction (but not in direct proportion).

The incident and refracted rays are reversible, that is, if the direction of propagation of the refracted ray in the second medium is reversed, the refracted ray in the first medium will propagate in the initial direction of the incident ray (Fig. 32.11a and b). (Prove it yourself.) Hence, when a ray passes from a medium of greater optical density to a medium of smaller optical density it deflects from the normal. Obviously in this case the value of the refractive index will be less than unity.

It is the deflection of the light rays from the normal caused by refraction which explains the apparent decrease in the depth of a water basin (Fig. 32.12a). A man sees at the depth  $h_1$  the virtual image  $K_1$  of a stone  $K$  actually lying at a depth  $h$ . (Prove that  $h/h_1 = n$ , where  $n$  is the refractive index of water with respect to air.)

When a man looks into water sideways, the stone appears to be displaced in the horizontal direction (towards the observer), since he sees the virtual image of the stone (Fig. 32.12b) whose position depends on the angle of incidence of the rays reaching his eye.

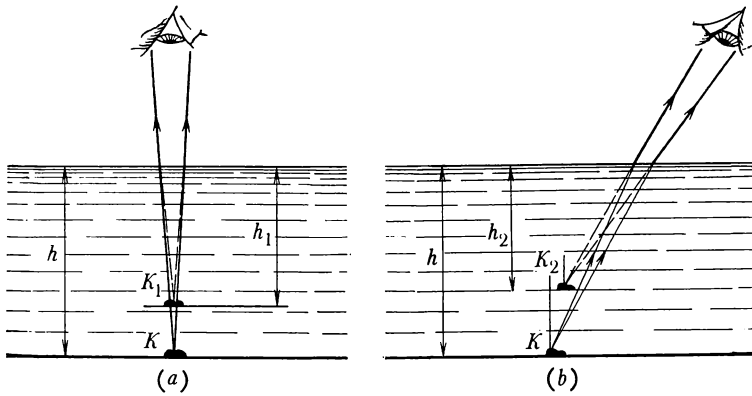


Fig. 32.12 As we look into water we see virtual images of objects in it and our estimate of their positions is erroneous.

Rays falling on the boundary surface separating transparent media at a normal enter the second medium without refraction.

### 32-5 Absolute and Relative Refractive Indices

The refractive index for rays entering a medium out of a vacuum is termed *absolute* (see Section 30-6) and is computed with the aid of formula (30.10):

$$n = c/v$$

In calculations, absolute refractive indices should be taken from tables. Since  $c$  exceeds  $v$ , the absolute refractive index always exceeds unity.

The formula for the second law of refraction when radiation enters a medium from a vacuum is

$$\frac{\sin i}{\sin \beta} = n \quad (32.5)$$

Formula (32.5) is often used also in cases when one of the media is air, since the propagation velocity of light in air is quite close to  $c$ , the absolute refractive index for air being 1.0029.

The formula for the second law of refraction when radiation enters a vacuum (air) from a medium is

$$\frac{\sin i}{\sin \beta} = \frac{1}{n} \quad (32.6)$$

In this case the rays leaving a medium always move away from the normal towards the boundary between the medium

and vacuum. Let us now find the method for calculating relative refractive indices  $n_{21}$  from the absolute refractive indices.

Let light enter into a medium with the absolute refractive index  $n_2$  from a medium with the absolute refractive index  $n_1$ . We may write

$$n_1 = \frac{c}{v_1} \quad \text{and} \quad n_2 = \frac{c}{v_2}$$

whence

$$\frac{n_2}{n_1} = \frac{v_1}{v_2} = n_{21}$$

Therefore the relative refractive index for rays entering the second medium out of the first is determined by the relation

$$n_{21} = \frac{n_2}{n_1} \quad (32.7)$$

The formula for the second law of refraction in this case is often written in the form

$$\frac{\sin i}{\sin \beta} = \frac{n_2}{n_1} \quad (32.8)$$

We recall that according to Maxwell's theory the absolute refractive index can be found from relation (30.14):

$$n = \sqrt{\mu \epsilon}$$

Since for substances transparent to light  $\mu$  is practically equal to unity, it can be assumed that

$$n = \sqrt{\epsilon} \quad (32.9)$$

**Table 32.2** Electrostatic ( $\epsilon_{st}$ ) and optical ( $\epsilon = n^2$ ) permittivities for some substances

Substance	$\epsilon_{st}$	$\epsilon = n^2$
Diamond	5.7	5.85
Naphthalene	2.5	2.5
Polystyrene	2.6	2.4
Sulphur	3.85	3.9
Ice	94	1.71
KCl	4.68	2.43
NaCl	5.6	2.25
NaF	6.0	1.74
Water	81	1.77

Since the frequency of luminous radiation is of the order of  $10^{14}$  Hz, neither dipoles nor ions of the dielectric, due to their large masses, are able to follow this frequency. The dielectric properties of a substance at such frequencies are determined mainly by the electronic polarization of its atoms (see Section 17.10). This explains the difference between  $\epsilon$  in (32.9) and in electrostatics. In an atomic substance made up of atoms of a single type and not containing any ions or natural dipoles the polarization can only be electronic. For such substances the values of  $\epsilon$  in the optical range and in electrostatics coincide. Diamond, which consists of carbon atoms, is an example of such a substance. For ionic solid dielectrics the electrostatic value  $\epsilon_{st}$  is much higher than the optical,  $\epsilon$  (32.9) (see Table 32.2). The difference between  $\epsilon_{st}$  and  $\epsilon$  reflects the contribution of ionic polarization to the permittivity of such a dielectric.

We recall that sea water is a good conductor of electricity, but a poor absorber of visible radiation, sharply differing from metals in this respect. This property of sea water is due to the absence of free electrons in it and the presence of a comparatively large number of ions unable to follow high-frequency oscillations of the light ray.

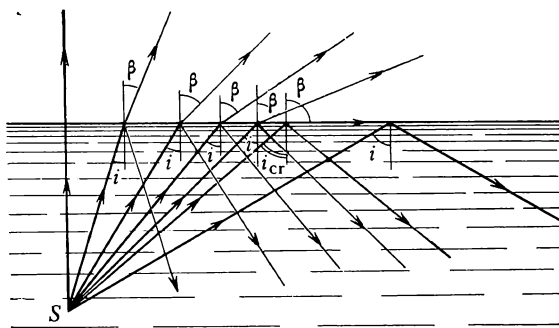
Note that the absolute refractive index depends not only on the substance but also on the frequency (or wavelength) of radiation. As a rule the refractive index increases with a decrease in wavelength (see Table 32.3).

**Table 32.3** Absolute refractive indices for water and glass at various wavelengths

Substance	Wavelength, $\mu\text{m}$				
	0.759	0.687	0.589	0.486	0.397
Glass (crown)	1.510	1.512	1.515	1.521	1.531
Water	1.329	1.331	1.333	1.337	1.344

### 32-6 Total Reflection

Place a light source in some transparent medium and observe the passage of light into a medium of smaller optical density, for instance air (Fig. 32.13).



**Fig. 32.13** Total reflection.

At the boundary surface the light will be both refracted and reflected; the energy of the reflected light will increase as the angle of incidence is increased and the energy of the refracted light will diminish. At an angle of incidence  $i_{cr}$  the refracted angle can be seen to glide along the boundary surface, while at angles greater than  $i_{cr}$  there are no refracted

rays at all. Such a phenomenon is observed only in cases when the incident light propagates in a medium of greater optical density, that is, when the refracted rays move away from the normal to the boundary surface separating the two media. The term for the phenomenon in which light is totally reflected from the boundary surface separating two transparent media is *total reflection*.

When the angle of incidence becomes equal to  $i_{cr}$ , the partially reflected rays become totally reflected (see Fig. 32.13). The term for the angle of incidence  $i_{cr}$  for which the angle of refraction  $\beta$  is equal to  $\pi/2$  is *critical angle*. Note that only rays falling on the boundary surface at angles greater than  $i_{cr}$  are totally reflected. The value of the critical angle can in all cases be determined from the relative refractive index of the media. Indeed, since for the angle  $i_{cr}$  the angle  $\beta = \pi/2$ , it follows from formula (32.8) that

$$\frac{\sin i_{cr}}{\sin (\pi/2)} = \frac{n_2}{n_1}$$

Since  $\sin (\pi/2) = 1$ , we finally obtain

$$\sin i_{cr} = \frac{n_2}{n_1} \quad (32.10)$$

When luminous radiation enters a vacuum out of some medium, (32.10) assumes the form

$$\sin i_{cr} = \frac{1}{n} \quad (32.11)$$

(Explain why the total reflection of rays emerging from a medium of smaller optical density and entering a medium of greater optical density is impossible.)

An empty test tube and gas bubbles in water sometimes shine like silver. This phenomenon is explained by the total reflection of rays at the boundary between water or glass with a gaseous medium. Total reflection is used for making optical guide fibers. The light is directed inside a transparent fiber from one end and emerges from the other. On its way it undergoes many total reflections from the fiber walls and so follows all the bends of the fiber.

### 32-7 Refraction by a Plane Parallel Plate and a Prism

Let us examine how a transparent plate with two plane parallel faces changes the path of light rays. Good quality window glass is an example of such a plate.

Let a narrow beam of light  $AO_1$  fall on a plate made of a substance with a refractive index  $n$  at angle  $i_1$  (Fig. 32.14). After refraction at the upper face the beam travels inside the plate along the path  $O_1O_2$ , is refracted again at the lower face and travels through the air along the path  $O_2B$ . Compare the angles  $i_1$  and  $\beta_2$ . The formula for the second law of refraction assumes for the upper face the form

$$\frac{\sin i_1}{\sin \beta_1} = n$$

and for the lower face

$$\frac{\sin i_2}{\sin \beta_2} = \frac{1}{n}$$

Since  $\angle \beta_1$  and  $\angle i_2$  are equal, we obtain after multiplying the equations term by term

$$\frac{\sin i_1}{\sin \beta_2} = 1$$

whence  $\sin i_1 = \sin \beta_2$  and  $\angle i_1 = \angle \beta_2$ . This means that the ray  $AO_1$  is parallel to the ray  $O_2B$ . Therefore light rays passing through a plate with plane parallel faces are displaced parallel to themselves. The thicker the plate and the greater the refractive index of its material the greater the displacement. The displacement  $d$  depends in addition on the angle of incidence  $i_1$ . Therefore when a person looks at objects through a thick transparent plate, he sees all objects displaced (Fig. 32.15).

Applied optics often makes use of triangular prisms. The two faces of the prism through which rays enter and emerge are termed *refraction faces*, the term for the angle  $\varphi$  between these faces being *refraction angle* of the prism (Fig. 32.16).

Let a narrow beam of light  $AO_1$  of a specified colour fall from the air on the prism with a refraction index  $n$ . Inside the prism it follows the path  $O_1O_2$ . Emerging from the prism it moves away from the normal to the face and follows the path  $O_2B$ . Hence the effect of the prism is to deflect the beam towards its base. Since the direction of the light before it entered the prism was  $AO_1$  and after it left it became  $O_2B$ , the beam was deflected by the angle  $\delta$  (see Fig. 32.16a), termed *the angle of deflection*. This angle is the greater the greater the refractive index of the material of the prism or its refraction angle  $\varphi$ .

If a triangular prism is placed in a medium for which its relative refractive index is below unity, the ray  $AO_1$  (see Fig. 32.16b), after passing through the prism, will be deflected by the angle  $\delta$  not towards the base but towards the

Fig. 32.14 Refraction by plane plate with parallel sides.

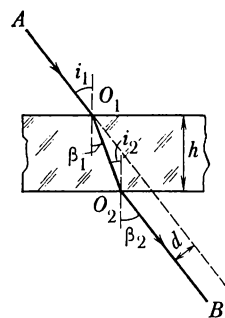


Fig. 32.15 Object seen through plate with parallel faces appears displaced.

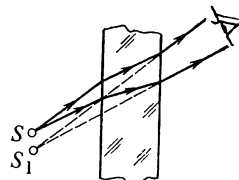
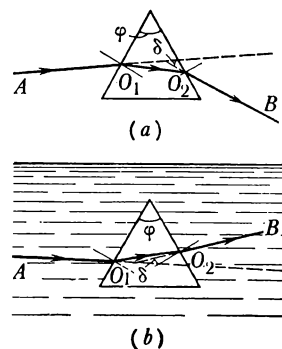


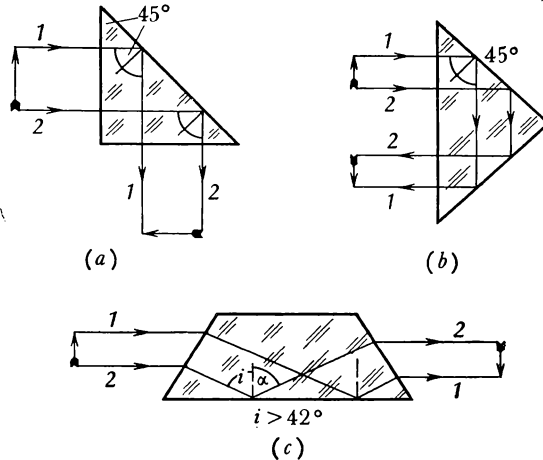
Fig. 32.16 Path of monochromatic ray through triangular prism made of material of (a) greater optical density than medium and (b) lesser optical density.



apex of the prism. (Explain why parallel rays remain parallel on emerging from the prism.)

Calculation shows the critical angle for glass to be about  $42^\circ$ . Therefore it is easy to obtain total reflection in a triangular glass prism with two angles of  $45^\circ$ . Figure 32.17*a* shows a change in direction of light rays of  $90^\circ$  in such a prism, while Fig. 32.17*b* shows the reversal of an image by

Fig. 32.17 Path of light rays in total reflection prisms.



the same prism. Figure 32.17*c* shows a direct-vision prism and the path of rays in it. The upper and lower rays are seen to change places, but they continue to propagate in the original direction.

There is one more point to be noted. Since the index of refraction  $n$  depends on the wavelength  $\lambda$ , the deflection of the rays in a prism also depends on their colour. For instance, the angle of deflection,  $\delta$ , for red rays is less than for blue. This problem is discussed at length in Section 37-2.

## 33

# Image Formation by Spherical Lenses and Mirrors

### 33-1 Lenses

A device frequently used to obtain images in optical instruments is the lens.

A *lens* is a transparent body bounded by two smooth convex or concave surfaces (one of them can be a plane).

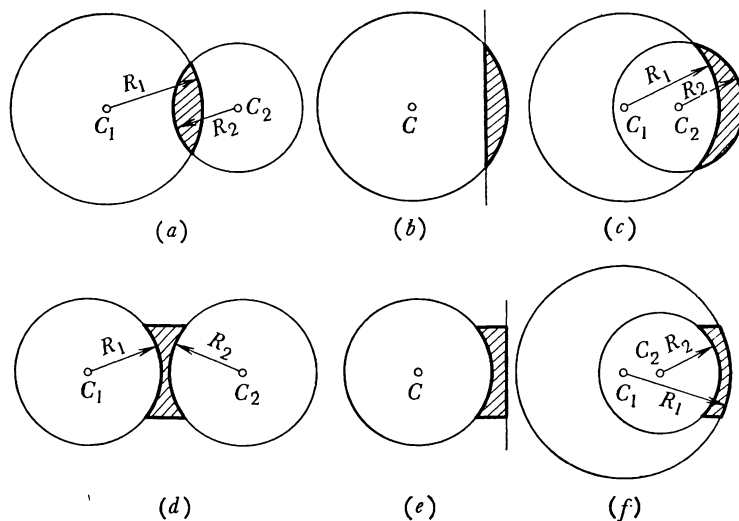


Fig. 33.1 Spherical lenses.

Most common are spherical lenses made of special kinds of glass, for instance of flint, or of other materials with an appropriate refractive index. The lenses can be of the convex type (Fig. 33.1a, b and c), that is, thicker in the middle, or of the concave type (Fig. 33.1d, e and f), that is, thinner in the middle.

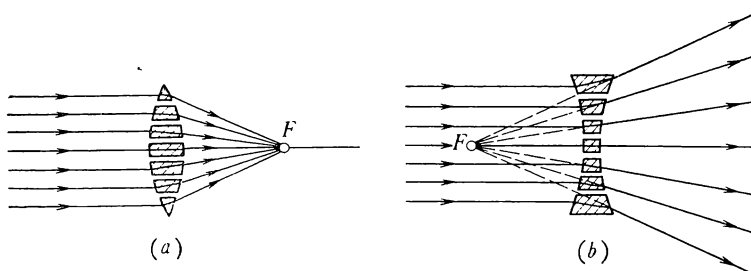


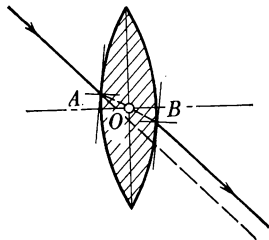
Fig. 33.2 Lens models assembled from prisms with different angles of refraction: (a) convex lens; (b) concave lens.

The straight line passing through the centres of curvature of the lens surfaces  $C_1$  and  $C_2$  or through the centre  $C$  and normal to the plane surface of the lens is termed *principal optical axis* of the lens. A ray of light travelling along the optical axis passes through the lens without refraction. (Why?)

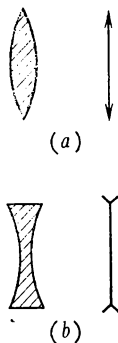
The changes in the path of the rays introduced by a lens can easily be established with the aid of a model made of prisms (Fig. 33.2). The prisms can be chosen so that parallel rays after passing through them will all meet at one point  $F$  (Fig. 33.2a). When the prisms are drawn together they



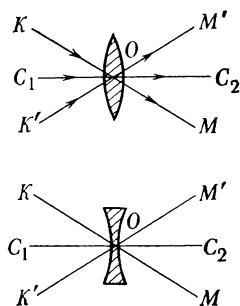
**Fig. 33.3** Ray passing through centre of lens does not change its direction.



**Fig. 33.4** (a) Converging and (b) diverging lenses and their symbols.



**Fig. 33.5**  $O$ , optical centre of lens;  $C_1C_2$ , principal optical axis;  $KM$  and  $K'M'$ , secondary optical axes.



form a body resembling a convex lens in shape. A convex lens possesses the property of collecting parallel rays at one point. For this reason convex lenses are termed *converging*. A model of a concave lens is shown in Fig. 33.2b. (Explain why concave lenses are termed *diverging*.)

Each lens has a point  $O$  on its principal optical axis (Fig. 33.3) with the remarkable property that a ray passing through it travels in the direction it travelled before entering the lens. The term for point  $O$  is *optical centre* of the lens. The planes tangent to the surfaces of the lens at points  $A$  and  $B$  are parallel, so the ray passing through point  $O$  passes through the lens as through a plane parallel plate, that is, it is displaced parallel to itself without a change in direction. Since this displacement is the smaller the thinner the plate, in sufficiently thin plates it can be neglected, especially if the angle the ray makes with the principal optical axis of the lens is small. In what follows we shall discuss only thin lenses of small diameter operating with rays making only small angles with the principal optical axis of the lens. Thin lenses are depicted schematically in Fig. 33.4. A ray passing through the optical centre of a thin lens can be assumed to pass through without refraction. Any straight line passing through the optical centre of a lens,  $O$ , (except the principal optical axis) is called a *secondary optical axis* (Fig. 33.5).

### 33-2 Focal Points and Planes

If a beam of rays parallel to the principal optical axis of a converging lens is directed at it, the rays will converge to a point  $F$  on the other side of the lens (Fig. 33.6a). Such rays emerging from a diverging lens will diverge (Fig. 33.6b), but in such a manner that their continuations will meet at a point  $F$  this side of the lens.

The term for the point  $F$  on the principal optical axis of the lens in which the rays originally parallel to the principal optical axis of the lens assemble is *principal focus* of the lens. It follows from the above that the principal focus of a converging lens is real and that of a diverging lens virtual. Every lens has two principal foci symmetrical about its optical centre  $O$ . The term for the distance  $f$  between the principal focus of a lens and its optical centre is *principal focal length*. For a real principal focus this distance is assumed to be positive and for a virtual one negative.

When rays fall on a lens parallel to its secondary axis, for instance  $AO$  (Fig. 33.7), after refraction they meet at one point  $B$  lying on the same axis. The term for this point is a *focus* of a lens. Obviously a lens may have an infinite number of various foci, but experiment shows that all of them lie in a *focal plane*,  $KM$ . The focal plane is a plane passing through the principal focus at right angles to the principal optical axis. Every lens has two focal planes.

Thus, rays parallel to any optical axis of a lens converge after refraction to the point of intersection of this axis with the focal point of the lens (see Fig. 33.7a). A converging lens has real and a diverging lens virtual focal planes (see Fig. 33.7b).

### 33-3 Lens Power

The position of the principal focus greatly affects the dimension and the shape of images obtained with its aid.

The quantity  $P$  characterizing the optical properties of the lens determined by the position of its principal focus on its optical axis is termed *lens power*. The measure of the lens power is the number reciprocal to its principal focus length  $f$ :

$$P = 1/f \quad (33.1)$$

The unit of measurement of lens power follows from (33.1):

$$P = 1/1 \text{ m} = 1 \text{ m}^{-1} = 1 \text{ D (diopter)}$$

The unit for measuring lens power in the SI system is the *diopter*—the power of a lens whose principal focal length is equal to one meter. It has been generally agreed to take the lens power of converging lenses (and thus focal length  $f$ ) to be positive and that of diverging lenses to be negative.

The power of a lens is determined by the curvature of its surfaces and by the refractive index of its material with respect to the medium it is acting in; it can be calculated with the aid of the formula

$$P = (n - 1) \left( \frac{1}{R_1} + \frac{1}{R_2} \right) \quad (33.2)$$

where  $R_1$  and  $R_2$  are the radii of the spherical surfaces of the lens and  $n$  is the refractive index relative to the medium the lens is acting in. When calculating the numerical value of  $R$  for a convex lens surface should be assumed to be positive and for the concave negative. Note that in conditions when

Fig. 33.6 Principal foci of (a) converging and (b) diverging lenses.

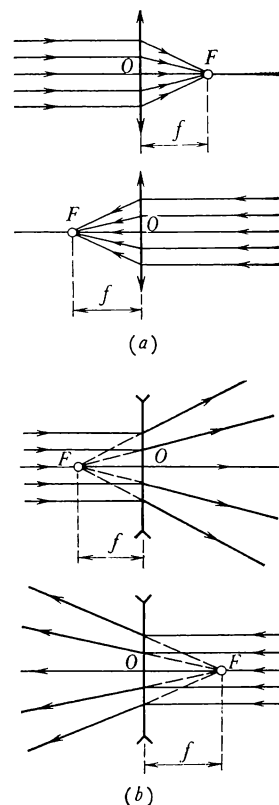
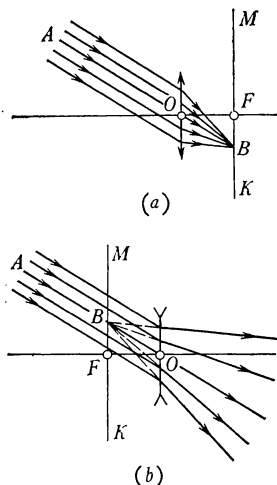


Fig. 33.7 All foci of lens lie in focal plane,  $KM$ .



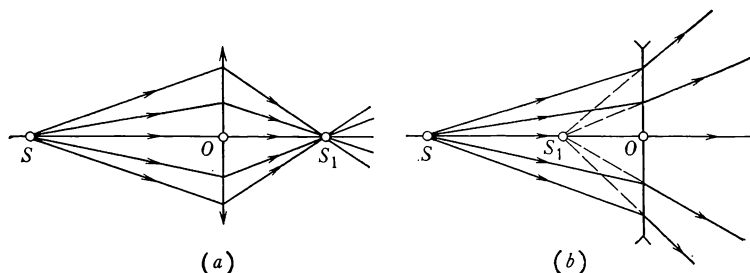
$n < 1$ , that is, when the lens is made of a material of smaller optical density than that of the medium, convex lenses act as diverging lenses and concave as converging.

### 33-4 Image Formation for a Luminous Point Lying on the Principal Axis of a Lens

A lens can be used to converge not only parallel rays on a point. Experiment shows that rays falling on a lens from a point  $S$ , after passing through the lens also converge on a point  $S_1$  (Fig. 33.8a), that is, the lens creates a real image of the luminous point  $S$  at point  $S_1$ . An image may also be virtual. Figure 33.8b depicts the path of rays falling from point  $S$  on a diverging lens. The rays, after passing through the lens, travel in a diverging beam, but in such a way that their continuation in the opposite direction converges at point  $S_1$ . Let us consider the method of constructing the lens image of a luminous point lying on the principal optical axis.

*First case:* point  $S$  is behind the principal focus of the lens (Fig. 33.9). Since all rays refracted by the lens converge at point  $S_1$ , it will suffice to determine where two such rays will intersect.

Fig. 33.8 (a) Point  $S_1$ , real image of point  $S$ ; (b) point  $S_1$ , virtual image of point  $S$ .



Let the straight line  $FO$  be the principal optical axis of a converging lens and  $KM$ , the focal plane of this lens. The ray passing from point  $S$  along the principal optical axis passes through the lens without refraction; therefore, the image of point  $S$  will lie on the principal optical axis  $FO$ . To find the exact place of the image of the point  $S$  we find the path of an arbitrary ray  $SA$  behind the lens. To this end draw an optical axis parallel to the ray  $SA$ . Suppose it intersects the focal plane  $KM$  at point  $A_1$ . The straight line joining points  $A$  and  $A_1$  is the path of the ray  $SA$  after its refraction in the lens. Continuing the straight line  $AA_1$  until it intersects the principal optical axis, we obtain

point  $S_1$ , which determines the position of the image of point  $S$  created by the lens. It is quite obvious that any other ray  $SB$  after refraction in the lens will also arrive at point  $S_1$  (see Fig. 33.9); the secondary optical axis  $OB_1$  is parallel to the ray  $SB$ .

*Second case:* point  $S$  lies between the principal focus and the optical centre of the lens (Fig. 33.10). As in the first case the image of point  $S$  will be at some point on the principal optical axis. To find where exactly, we choose an arbitrary ray  $SA$  falling on the lens. Draw a secondary optical axis  $OA_1$  parallel to  $SA$  and then the straight line  $AA_1$  until it intersects the optical axis at point  $S_1$ . The latter gives us the position of the virtual image of  $S$  in the case being considered.

*Third case:* the bright point is on the principal optical axis of a diverging lens (Fig. 33.11). In this case when constructing an image the focal plane should be arranged on the same side of the lens as point  $S$ . In this case, too, the image of the bright point  $S$  should lie on the principal optical axis. Choose an arbitrary ray  $SA$  and draw a secondary axis  $OA_1$  parallel to it. The point of intersection of the straight line  $AA_1$  with the principal optical axis will determine the position of the virtual image  $S_1$ . Note that the image of a real luminous point produced by a diverging lens is always virtual.

### 33-5 The Lens Formula

In the preceding section it was explained that the position of the image  $S_1$  of a luminous point  $S$  is always uniquely determined by the position of the bright point itself with respect to the lens. For this reason the points  $S$  and  $S_1$  are termed *conjugate points* of the lens. Let us deduce a formula for conjugate points of a lens which would enable us to find the position of the image  $S_1$  with the aid of calculations.

Let there be a bright point  $S$  lying on the principal optical axis of a converging lens with the centre at  $O$  and the foci at  $F_1$  and  $F_2$  (Fig. 33.12), and let its image be at point  $S_1$ . Recall that  $KM$  is the focal plane of the lens and that  $OA_1 \parallel SA$ . Denote the distance from the bright point  $S$  to the optical centre  $O$  by  $d$  ( $OS = d$ ), the distance from the image  $S_1$  to the optical centre  $O$  by  $s$  ( $OS_1 = s$ ), and the principal focal length by  $f$  ( $OF_1 = f$ ). We obtain from the condition of similarity of triangles  $SAS_1$  and  $OA_1S_1$  ( $OA_1 \parallel SA$ )

$$SS_1/OS_1 = AS_1/A_1S_1, \text{ or } (d+s)/s = AS_1/A_1S_1$$

Fig. 33.9 Image formation of luminous point  $S$  lying on principal optical axis of converging lens behind principal focus,  $F$ .

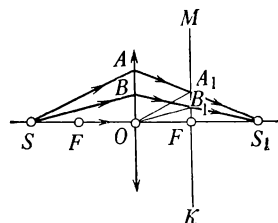


Fig. 33.10 Image formation of luminous point  $S$  lying on principal optical axis between principal focus and optical centre of converging lens.

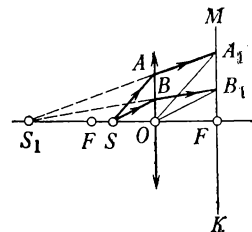
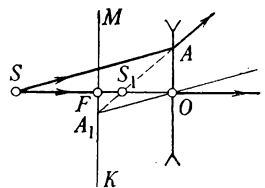


Fig. 33.11 Image formation of luminous point  $S$  lying on principal optical axis of diverging lens.



From the condition of similarity of triangles  $OAS_1$  and  $F_1A_1S_1$  we may write

$$OS_1/F_1S_1 = AS_1/A_1S_1, \text{ or } s(s-f) = AS_1/A_1S_1$$

Since the right-hand sides of the above proportions are equal, we have

$$(d+s)/s = s/(s-f)$$

whence

$$sf + df = ds$$

Dividing both sides of the equation by  $dsf$ , we obtain the formula for the conjugate points of a lens:

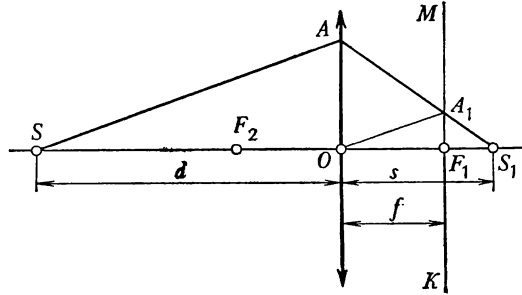
$$1/d + 1/s = 1/f \quad (33.3)$$

Since the term on the right-hand side is the lens power, we obtain

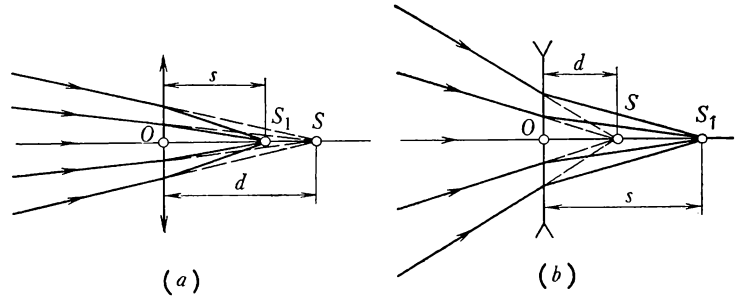
$$1/d + 1/s = P \quad (33.4)$$

It follows from (33.3) that the formula remains true if  $d$  and  $s$  change places. This means that a luminous point and

**Fig. 33.12** Conjugate points of lens.



**Fig. 33.13** Virtual light source and its real image in (a) converging and (b) diverging lenses.



its image in a lens are interchangeable, that is, if the luminous point is placed in the location of the image, the new image will be in the former location of the point. This is why points  $S$  and  $S_1$  are termed *conjugate*.

One should keep in mind that relations (33.3) and (33.4) are valid both for converging and diverging lenses. In calculations the values of real quantities are always substituted into the formulae with a plus sign and the values of virtual quantities with a minus sign. For instance, when relation (33.3) is used for a diverging lens, a number with a minus sign is substituted in place of  $f$ . A negative result obtained in the calculation means that the quantity corresponding to it is virtual.

Recall that the luminous point  $S$ , too, can be virtual. Figure 33.13a depicts a virtual light source  $S$  and its real image  $S_1$  in a converging lens, and Fig. 33.13b a virtual light source  $S$  and its real image  $S_1$  in a diverging lens.

### 33-6 Image Formation for a Luminous Point Lying on a Secondary Axis of a Lens

The image of a luminous point  $S$  lying on a secondary optical axis of a lens lies on the same axis. Let us find out how the image should be constructed.

*First case:* point  $S$  is behind the focal plane of a converging lens (Fig. 33.14). Any two of the three rays depicted in

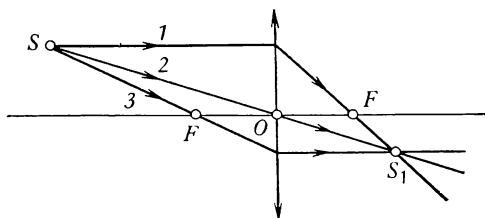
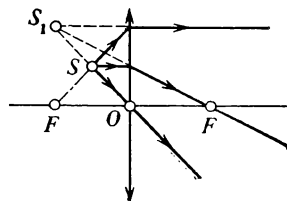


Fig. 33.14 Image formation of luminous point  $S$  lying off principal optical axis of converging lens at distance from lens exceeding  $f$ .

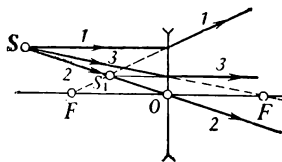
Fig. 33.14 can be used to locate the image  $S_1$ . The ray 1 is drawn from point  $S$  parallel to the principal optical axis. After being refracted in the lens it passes through the principal focus. The ray 2 follows the secondary axis, that is, it passes through the optical centre of the lens. This ray passes through the lens without refraction. The ray 3 is drawn through the principal focus  $F$ . After being refracted in the lens it follows a path parallel to the principal optical axis. The point of intersection of all these refracted rays  $S_1$  gives us the location of the real image of point  $S$  in this case.

*Second case:* point  $S$  is between the focal plane of a converging lens and the lens itself (Fig. 33.15). In this case, too, three rays can be drawn from point  $S$  perfectly similar to those in the first case. The point of intersection of any

Fig. 33.15 Image formation of luminous point  $S$  lying off principal optical axis of converging lens at distance from lens less than  $f$ .



**Fig. 33.16** Image formation of luminous point  $S$  lying off principal optical axis of diverging lens.



two of them,  $S_1$ , gives us the location of the virtual image of  $S$  in the case discussed.

*Third case:* point  $S$  lies on the secondary optical axis of a diverging lens (Fig. 33.16). In this case, too, three rays similar to those in the first case can be drawn from point  $S$ . However, it should be kept in mind that after refraction in the lens the continuation of ray 1 must pass through the focus that lies on the same side of the lens as point  $S$ . Ray 3 should be drawn so that its continuation passes through the focal point to the other side of the lens. Then, after refraction in the lens, the ray will take a path parallel to the principal optical axis. Note that the image of a real luminous point  $S$  in a diverging lens is always virtual.

### 33-7 Image Formation by Spherical Lenses

Let there be an object in front of the lens which in the following will be designated by an arrow at right angles to the principal optical axis. The image of the object formed by the lens is the sum of the images of its individual points. Therefore, to construct its image one needs only to find the locations of the images of its extreme points.

Various typical cases of constructing images of the object  $AB$  using a converging lens are illustrated in Fig. 33.17. The method of construction is as follows. First the image of point  $A$  is constructed, then that of point  $B$ . The points  $A_1$  and  $B_1$  thus obtained are joined by a straight line  $A_1B_1$ . This becomes the image of  $AB$ .

*First case:* the distance  $d$  from the object to the lens exceeds  $2f$  (Fig. 33.17a). In this case the object and its image in the lens are on opposite sides of the lens and the distance from the image to the lens,  $s$ , exceeds  $f$ , but is less than  $2f$ . The image itself will be real, inverted and reduced in size. Specifically, when a luminous object is an infinite distance away from the lens ( $d = \infty$ ), its image will take the form of a luminous point located at the principal focus of the lens (Fig. 33.17b).

*Second case:* the distance from the object to the lens  $d$  is  $2f$  (Fig. 33.17c). In this case the object and its image are on opposite sides of the lens and the distance from the lens to the image is  $s = d = 2f$ . The image itself will be real, inverted and full-scale.

*Third case:* the distance from the object to the lens  $d$  is greater than  $f$  but less than  $2f$  (Fig. 33.17d). In this case the object and its image in the lens are on opposite sides of the

lens and the distance from the lens to the image exceeds  $2f$ . The image itself will be real, inverted and magnified.

*Fourth case:* the object is at the principal focus of the lens, that is, the distance of the object from the lens is  $d = f$  (Fig. 33.17e). In this case the rays from every point of the

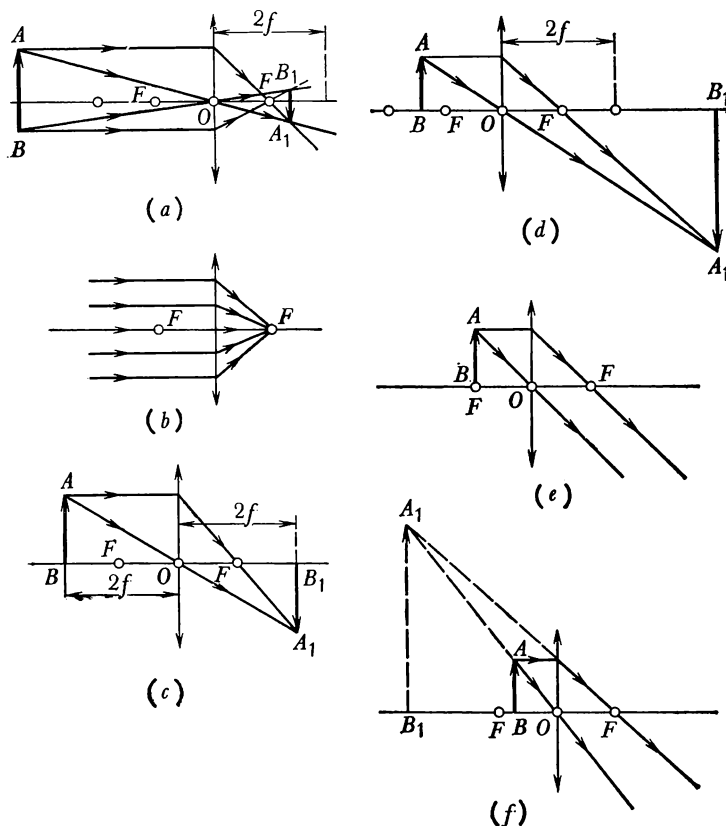


Fig. 33.17 Image formation of objects in different positions with respect to converging lens.

object after refraction in the lens will travel in a parallel beam. This means that the image must be of an infinite size and located at an infinite distance from the lens, that is, that practically there should be no image at all.

*Fifth case:* the distance from the lens  $d$  is less than the principal focal length  $f$  (Fig. 33.17f). In this case the object and its image are on one side of the lens and the distance from the lens to the image  $s$  exceeds  $d$ . The image itself is virtual, direct and magnified.

Let us observe the change in the image of the object when the latter is brought closer to the lens from infinity.



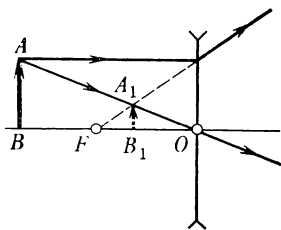
As the object is brought closer to the lens to within the distance  $2f$  from it its inverted real image travels a distance from  $f$  to  $2f$  gradually increasing in size but never reaching the full size of the object. With the object at a distance of  $2f$  from the lens its inverted image is at the same distance  $2f$  from the lens. As the object is brought still closer to the lens to within the distance  $f$ , its image, which now exceeds the object in size, continues to increase and moves to infinity.

Lastly, when the object moves from the principal focus towards the lens its virtual image, which is now behind the object, diminishes in size and draws closer to the lens. When the object is in contact with the lens its virtual image coincides in size and in location with the object.

Note that the transition of the object from the opposite side of the lens to the near side takes place at the moment the object crosses the focal point of the lens. Hence, the object and its image always move in one direction.

The construction of the image of an object formed by a divergent lens is illustrated in Fig. 33.18. A divergent lens always forms a virtual direct image of an object, and the image is reduced in size and located between the principal focus and the lens. The distance from this image to the lens  $s$  is always less than that from the object to the lens  $d$ . In this case, too, the object and its image always move in the same direction, the image coinciding with the object as the latter makes contact with the lens.

Fig. 33.18 Image formation by diverging lens.



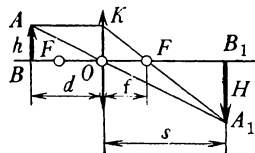
### 33-8 Lateral Magnification

It was established in the preceding section that a lens can be used to produce magnified images of objects. They are frequently so used in practice.

The term *lateral magnification*,  $\beta$ , applies to the ratio of the height (width) of the image of an object to its actual height (width). If the height of the object is denoted by  $h$  and that of its image by  $H$ ,

$$\beta = H/h \quad (33.5)$$

Fig. 33.19 Lateral magnification of lens.



Let us find the dependence of  $\beta$  on the distances from the lens to the object  $d$  and to its image  $s$ . Figure 33.19 depicts the positions of the object  $AB$  and its image  $A_1B_1$  with respect to the lens. From the condition of similarity of  $\triangle AOB$  and  $\triangle A_1OB_1$  it follows that  $A_1B_1 \div AB = OB_1 \div OB$ . Since  $A_1B_1 = H$ ,  $AB = h$ ,  $OB_1 = s$  and  $OB = d$ , we obtain the formula for calculating the lateral magnification of

a lens:

$$H/h = s/d, \text{ or } \beta = s/d \quad (33.6)$$

It follows from these relations that a converging lens magnifies only if the image is farther from the lens than the object. (Can a diverging lens produce a magnification exceeding unity?)

### 33-9 Spherical Mirrors

Mirrors whose surface is a part of a sphere are termed spherical; they can be either concave or convex (Fig. 33.20). The diameter  $KM$  of the circumference bounding the mirror is termed *aperture* of the mirror, and the point  $O$  farthest from it is termed *apex* of the mirror. The term for the straight line passing through the centre of curvature of the mirror  $C$  and its apex  $O$  is *principal optical axis* of the mirror, and for any other straight line passing through point  $C$  and the surface of the mirror a *secondary optical axis* of the mirror.

When a ray of light follows an optical axis, its angle of incidence on the surface of the mirror is zero, and because of this such a ray travels after reflection along the same optical axis in the opposite direction. If a beam of rays travelling parallel to the principal optical axis of a mirror strikes its surface, after reflection the rays pass through point  $F$  lying on the principal optical axis (Fig. 33.21a), the point being termed the *principal focus* of the mirror. Hence, concave mirrors are *converging* mirrors. Their principal focus is a real one. The principal focus of convex mirrors is virtual (Fig. 33.21b). Such mirrors are *diverging* mirrors.

The term for the distance between the principal focus and the mirror's apex  $OF$  is *principal focal length*  $f$ . Let us find its relation to the mirror's radius of curvature,  $R$ .

The ray  $AA_1$  parallel to the principal optical axis of the mirror after reflection takes the path  $A_1F$  (Fig. 33.21a). Connect point  $A_1$  with the centre of curvature of the mirror,  $C$ . It follows from the laws of reflection that  $\angle 2 = \angle 3$ . Since  $A_1A \parallel OC$ , it follows that  $\angle 1 = \angle 2$ . Therefore  $\angle 1 = \angle 3$  and  $\triangle A_1FC$  is isosceles. Since the surface of a mirror is always only a small part of the spherical surface, it can be assumed that approximately  $A_1F \approx OF$ . Hence,  $FC \approx OF$ . This means that point  $F$  divides the mirror's radius  $OC$  in half:

$$f = R/2 \quad (33.7)$$

Fig. 33.20 (a) Concave spherical mirror; (b) convex spherical mirror;  $R$ , radius of curvature of mirror.

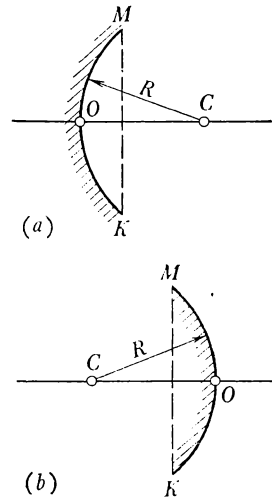
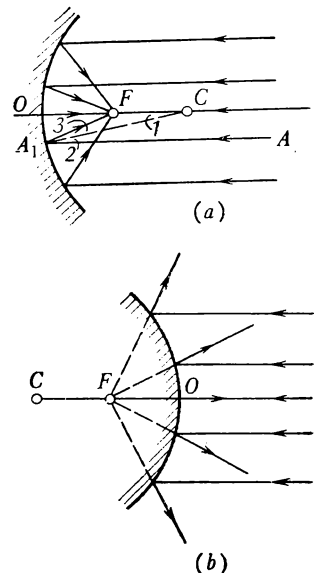
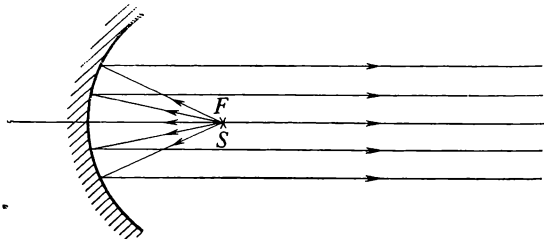


Fig. 33.21 Principal foci of (a) concave and (b) convex spherical mirrors.



As in the case of lenses the incident and reflected rays in spherical mirrors are reversible. Therefore, if a light source is placed at the principal focus of a concave mirror, the rays after reflection will travel practically parallel to the principal optical axis of the mirror (Fig. 33.22).

**Fig. 33.22** Formation of almost parallel bunch of light using concave spherical mirror.

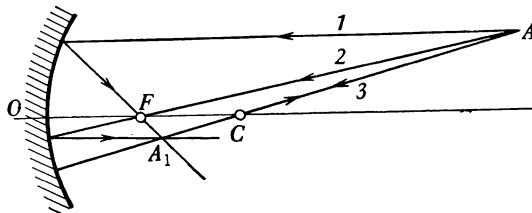


In practice parabolic mirrors, whose reflecting surface is part of a paraboloid of revolution, are used instead of spherical to produce parallel beams of light.\* Parabolic mirrors produce beams of superior directionality. This property of mirrors is utilized in the design of searchlights and reflectors of various types.

### 33-10 Image Formation by Spherical Mirrors

Mirrors, like lenses, can form various images of objects. To construct the image of a point  $A$  formed by a spherical mirror any two of the rays depicted in Fig. 33.23 can be used.

**Fig. 33.23** Image formation of point  $A$  by spherical mirror.



The ray 1 from point  $A$  is drawn parallel to the principal optical axis. After reflection it passes through the mirror's principal focus  $F$ . The ray 2 from point  $A$  is drawn from the principal focus  $F$ . After reflection it travels parallel to the principal optical axis. The ray 3 is drawn through the

\* A paraboloid of revolution is a geometric body obtained as the result of rotation of a parabola about its axis of symmetry.

centre of curvature. After reflection it travels back along the same straight line.

The typical cases of images formed by spherical mirrors are the same as of those formed by lenses. They are illustrated in Fig. 33.24. Note that a convex mirror always forms a virtual image. On the other hand, more objects can be seen from one point at a single time in a convex mirror than in a concave one. Convex mirrors are used by drivers to obtain a side and back view.

Formulae (33.3)-(33.6), previously deduced for lenses, are also valid for spherical mirrors. As before numerical values of real quantities should be substituted in these formulae with a plus, and those of virtual quantities with a minus.

Let us demonstrate how the formula for conjugate points of a spherical mirror is deduced. Let the point source  $A$  be on the principal optical axis of the mirror  $OC$  (Fig. 33.25). The ray  $AB$  after reflection passes through point  $A_1$ , the location of the image of point  $A$ . Denote the distance  $AO$  by  $d$ , the distance  $A_1O$  by  $s$  and  $OC$  by  $R$ . For mirrors whose surface is only a small part of a sphere it can be assumed that approximately  $BA \approx OA = d$  and  $BA_1 \approx OA_1 = s$ . The radius  $R$  drawn to the point of incidence  $B$  is normal to the mirror's surface; therefore  $\angle 1 = \angle 2$ . Consequently, the line  $BC$  in  $\triangle ABA_1$  is the bisector of the angle  $ABA_1$ . This means that the segments  $AC$  and  $A_1C$  are proportional to the sides of  $\triangle ABA_1$ :

$$A_1C/AC = BA_1/BA, \text{ or } (R - s)/(d - R) = s/d$$

Transform the latter relations thus:

$$Rd - sd = sd - R^2 \quad Rs + Rd = 2sd$$

Dividing by  $Rsd$ , we obtain

$$1/d + 1/s = 2/R \quad (33.8)$$

Substituting the value of  $R$  from (33.7), we obtain the formula for conjugate points of a mirror (33.3):

$$1/d + 1/s = 1/f$$

Demonstrate that formula (33.6) is valid for spherical mirrors.)

Fig. 33.24 Image formation of objects by spherical mirrors: (a) concave; (b) convex.

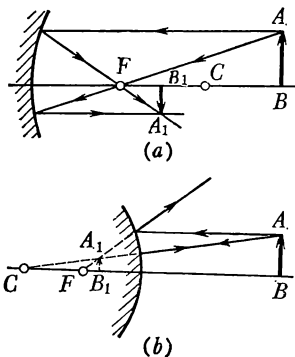
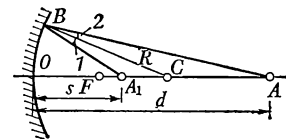


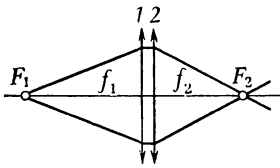
Fig. 33.25 Conjugate points of spherical mirror.



## 34 The Eye and Vision. Optical Instruments

### 34-1 Optical Systems

Fig. 34.1 Centred system of lenses in contact; foci of lenses are conjugate points.



Lenses have numerous deficiencies and rarely form a good image. To obtain good images frequent use is made of optical systems.

Several lenses with a common principal optical axis form a *centred optical system*. Sometimes the centred system is made up of lenses made of different materials and put together to eliminate deficiencies in the individual lenses. Let us consider such a system (Fig. 34.1).

If a light source is placed at the focus of the first lens  $F_1$  the rays refracted in it travel parallel to the principal optical axis and, emerging from the second lens, pass through its principal focus  $F_2$ . Hence, points  $F_1$  and  $F_2$  are conjugate for the system being considered; therefore we obtain from formula (33.3)

$$\frac{1}{f_1} + \frac{1}{f_2} = \frac{1}{f_s} \quad (34.1)$$

where  $f_s$  is the principal focal length of the system. Since  $1/f_1 = P_1$ ,  $1/f_2 = P_2$  and  $1/f_s = P_s$ , it follows that

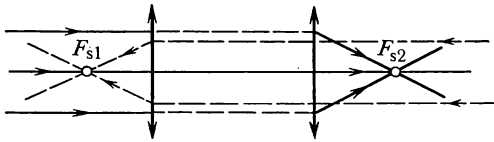
$$P_1 + P_2 = P_s \quad (34.1a)$$

Hence, the lens power of a system made up of thin lenses in contact is equal to the algebraic sum of the lens powers of the individual lenses.

Optical instruments often employ systems of two or more lenses placed at some distance from each other. To construct images formed by such systems one considers the successive effects of each lens, that is, one constructs an image of the object formed by the first lens, then regards it as the object for the second, and so on. However, this process, as we shall see, can be substantially simplified.

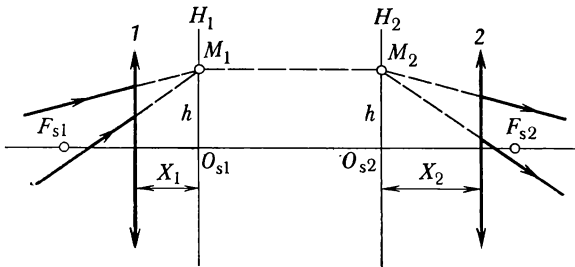
Let there be an optical system with a beam of rays parallel to its principal optical axis falling on it from the right (Fig. 34.2). Emerging from the system, the rays pass through point  $F_{s1}$ —the *principal focus of the system*. If rays parallel to the principal optical axis of the system fall on it from the left, on emerging from it they will pass through point  $F_{s2}$ —the second principal focus of the system.

It has been established that every system has two planes normal to its principal optical axis (planes  $H_1$  and  $H_2$  in Fig. 34.3) characterized by the following property. If a beam of rays is directed at the system from one side, for instance



**Fig. 34.2** Principal foci of optical system.

from the left, and these rays are expected to converge at point  $M_1$  in the plane  $H_1$ , on emerging from the system they will take the same paths as if their origin was point  $M_2$  in the plane  $H_2$ . So  $M_1O_{s1} = M_2O_{s2} = h$ . If the directions of the rays in Fig. 34.3 are reversed, the aforesaid will still be true. This means that if a luminous object is placed in plane  $H_1$ , its full-scale direct image will be in plane  $H_2$ , and vice versa. Thus, the points  $M_1$  and  $M_2$ , as well as the planes  $H_1$  and  $H_2$ , are conjugate.



**Fig. 34.3** Conjugate planes of optical system.

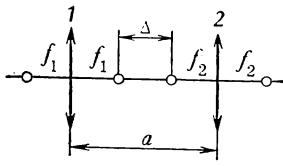
The term for conjugate planes  $H_1$  and  $H_2$  for which the lateral magnification is unity is *principal planes of the system*, and for the points of their intersection with the optical axis  $O_{s1}$  and  $O_{s2}$  *principal points of the system*. The focal lengths of the system  $O_{s1}F_{s1} = f_{s1}$  and  $O_{s2}F_{s2} = f_{s2}$  are measured from the principal points of the system  $O_{s1}$  and  $O_{s2}$ , the following sign rule being applied: a segment measured in the direction of the rays is assumed to be positive and in the direction opposite to that of the rays negative. Note that if the medium on both sides of the system is the same (this is the usual situation), the focal lengths of the system are equal in magnitude but opposite in sign:

$$f_{s1} = -f_{s2} \quad (34.2)$$

The expression for the lens power  $P_s$  of a system consisting of two thin lenses, of lens power  $P_1$  and  $P_2$ , placed in air at a distance  $a$  (Fig. 34.4) is

$$P_s = P_1 + P_2 - aP_1P_2 \quad (34.3)$$

Fig. 34.4 System of two lenses.



It follows from formula (34.3) that, by choosing the distance correctly, two converging lenses can be used as either a converging or a diverging system. If the distance  $a$  in (34.3) is made zero (the lenses are in contact), we obtain formula (34.1a). From the value of  $P_s$  found from (34.3) one can compute the focal length of the system using the formula  $P_s = 1/f_s$ .

Sometimes the position of the lenses in the system is described with the aid of a quantity  $\Delta$  instead of  $a$  (see Fig. 34.4), the relation between  $\Delta$  and  $a$  being

$$\Delta = a - f_1 - f_2 \quad (34.4)$$

In this case the numerical value of  $f_s$  is computed from the relation

$$f_s = -\frac{f_1 f_2}{\Delta} \quad (34.5)$$

Formula (34.5) can be easily obtained from (34.3) if one substitutes  $P_s = 1/f_s$ ,  $P_1 = 1/f_1$  and  $P_2 = 1/f_2$ .

The position of the principal planes of the system is determined by the following relations:

$$X_1 = a \frac{P_2}{P_s} \quad (34.6)$$

$$X_2 = -a \frac{P_1}{P_s} \quad (34.7)$$

where  $P_s$  is the system's lens power determined from formula (34.3). Here  $X_1$  is measured from the first lens and  $X_2$  from the second. Note that all sorts of variations in positions of the principal planes of a system are possible depending on the type of the system: they can be either inside the system or outside it, both on one side of the system or on opposite sides of it.

Knowing the position of the principal planes  $H_1$  and  $H_2$  and of the principal foci, we can construct images of objects without paying any attention to the path of the rays inside the system. For instance, to construct the image of point  $S_1$  one can use any two of the three rays shown in Fig. 34.5. The ray 1 is drawn parallel to the principal optical axis. Emerging from the plane  $H_2$  it passes through the focus  $F_{s2}$ . The ray 2 is drawn through the focus  $F_{s1}$ . Emerging from the plane  $H_2$  it proceeds parallel to the principal optical axis.

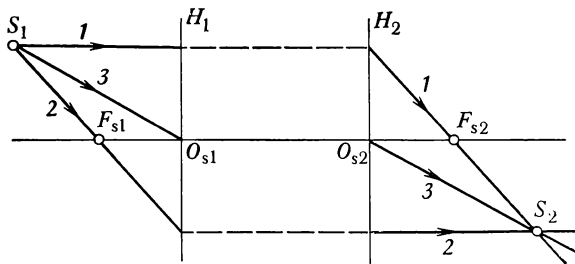


Fig. 34.5 Image formation of luminous point by optical system.

The ray 3 drawn from point  $S_1$  to point  $O_{s1}$ . Emerging from the plane  $H_2$  it proceeds from point  $O_{s2}$  parallel to  $S_1O_{s1}$ . Note that if we imagine the merging of planes  $H_1$  and  $H_2$ , we obtain the diagram used to construct the image of a point source shown in Fig. 33.14.

### 34-2 Deficiencies of Optical Systems

We now discuss the more important deficiencies of lenses and of optical systems.

The first deficiency is that rays emerging from a point  $S_1$  lying on the principal optical axis do not converge at one point (Fig. 34.6). The further the point of incidence of a ray

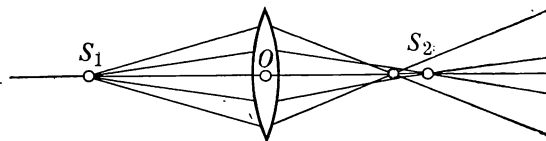


Fig. 34.6 Spherical aberration.

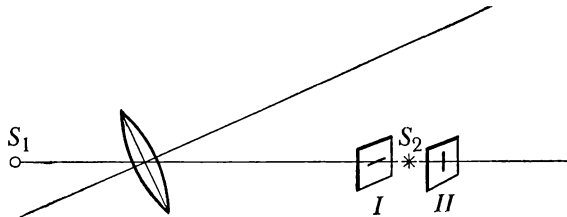
emanating from point  $S_1$  from the optical centre  $O$  the closer its point of intersection with the principal axis  $S_2$ . The term for this phenomenon is *spherical aberration* (from the Latin *aberratio* for deviation). This deficiency can be partially eliminated with the aid of a diaphragm which limits the width of the beam of rays reaching the lens. The diaphragm is made so that it can be used to vary the aperture through which light reaches the lens. Spherical aberration is also frequently corrected by joining two specially selected lenses. A complex lens or system free from spherical aberration is termed *aplanat*.

The second deficiency involves the image of a luminous point  $S_1$  located on a secondary optical axis (Fig. 34.7). Actually, in this case there will be two images in the form of



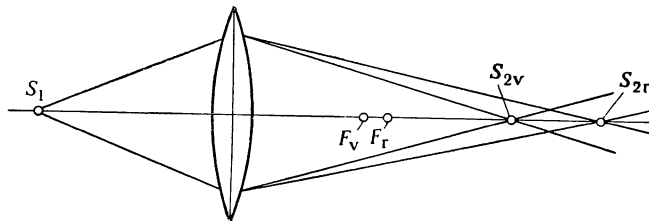
segments of straight lines arranged at right angles in the planes  $I$  and  $II$ . The image of  $S_1$  can be obtained only in the shape of a blurred illuminated circle  $S_2$  lying between the planes  $I$  and  $II$ . This phenomenon is termed *astigmatism*. The term for a lens or system free from this deficiency is *anastigmat*.

Fig. 34.7 Astigmatism.



It was stated in Section 32-5 that the refractive index depends on the frequency, that is, on the colour of the rays. It turns out that the focal point for violet rays,  $F_v$ , is nearer to the lens than that for red rays,  $F_r$  (Fig. 34.8). The

Fig. 34.8 Chromatic aberration: focus for violet rays is closer to lens than for red rays.



result is that the image of a white light point source is blurred and has coloured fringes. The term for this phenomenon is *chromatic aberration*. Lenses and optical systems in which this deficiency has been eliminated for two colours are termed *achromatic* and for three colours *apochromatic*.

### 34-3 Projection Lantern

The device used to show to an audience magnified images of transparent drawings or of drawings printed on nontransparent materials such as, for instance, paper, is the *projection lantern*. The term for the drawing made on glass or a transparent film and intended for showing on a screen is *slide*, and for the projector itself *diascope* (from the Greek *scopeo* for to see and *dia* for through).

The term for a projector designed for showing nontransparent drawings is *episcope* (from the Greek *epi* for on).

Projection lanterns which can be used to show both transparent and nontransparent pictures are termed *epidiascopes*. Let us see how a projection lantern works.

A lens which forms an image of an object in front of it is termed *objective*, or *projection lens*. Usually the objective is an optical system free of the most serious deficiencies to which individual lenses are subject. To make the image on the screen clearly visible to the audience the object itself must be brightly illuminated. This is achieved with the aid

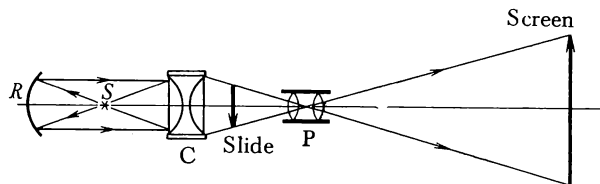


Fig. 34.9 Diagram of projection lantern.

of the device shown in Fig. 34.9, which depicts a schematic diagram of a projection lantern. The light source  $S$  is arranged at the focus of a concave mirror (reflector)  $R$ . The light travelling directly from the source  $S$  and reflected from the reflector  $R$  falls on the condenser  $C$  made of two plane-convex lenses. The condenser concentrates these light rays on the objective projection lens  $P$  which directs them onto the screen on which the image of the slide is demonstrated. The slide itself is placed between the principal focus of the objective and the point  $a$ , which is a distance of  $2f$  away from the objective. To obtain good definition the objective is adjusted, or *focused*.

### 34-4 The Photographic Camera

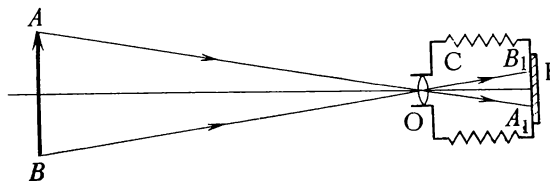
The term for an optical instrument used for taking photographs of objects is the *photographic camera*. It consists of light-proof chamber  $C$  (Fig. 34.10) with a mobile front wall housing the objective  $O$ .

When taking snapshots of the object  $AB$ , one obtains first a sharp image  $A_1B_1$  of the object on the back wall of the chamber by adjusting the objective. Next the objective is closed and a photographic plate, or film,  $F$  coated with a light-sensitive coating is placed in front of the back wall. Then the objective is opened for an interval of time termed *exposure*. During this interval chemical reactions take place in the light-sensitive layer forming the image of the object.

After development and fixing the image on the film becomes visible. On the image obtained the bright spots of the

object are dark and the dark spots are bright and transparent. For this reason such an image is termed *negative*. To obtain a normal photograph, termed *positive*, light-sensitive paper is placed on the negative and the latter is illuminated by light so that light falls on the paper through the negative.

Fig. 34.10 Diagram of photographic camera.



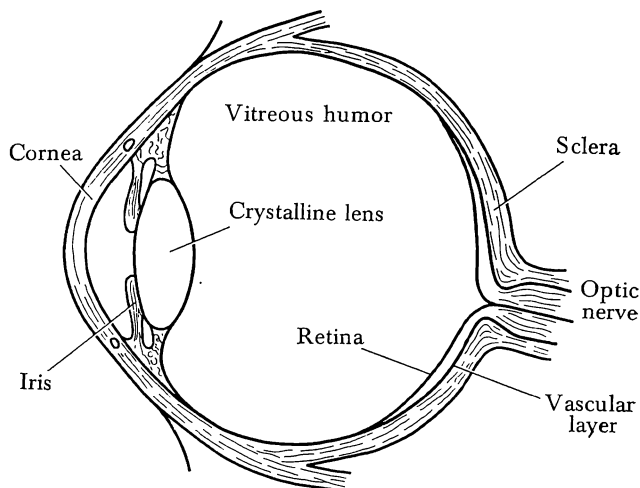
After some time a latent image is formed on the photographic paper. After the developing and fixing processes a normal photograph appears on the paper. One negative may be used to make many positives, or photographs.

### 34-5 The Eye as an Optical System

The human organ of sight is the eye, in many respects a perfect optical system. Let us see how this system works.

The eye is a spherical body of about 2.5 cm in diameter termed the *eyeball* (Fig. 34.11). The nontransparent and

Fig. 34.11 Diagram of the eye.



firm outer shell is called the *sclera*, while its transparent and more convex front part is the *cornea*. On the inside the sclera is covered with a *vascular layer* containing blood vessels feed-

ing the eye. The vascular layer opposite the cornea turns into the iris (different people have irises of different colour), separated from the cornea by a chamber containing a transparent watery fluid.

The iris has a circular orifice, termed *pupil*, of variable diameter. Thus the iris plays the part of a diaphragm controlling the amount of light entering the eye. Bright illumination makes the pupil contract, poor illumination makes it wider. Behind the iris inside the eyeball there is the *crystalline lens*—a double convex lens of a transparent material with a refractive index of about 1.4. Note that the radius of curvature of the internal surface of the crystalline lens is less than that of the external surface adjoining the iris. The crystalline lens is encircled by the *ciliary muscle*, which can change the curvature of its surfaces and, by force of this, its lens power.

The vascular layer on the inner surface of the eye carries branches of the optic nerve, the density of the branches being especially great opposite the pupil. These branches form the *retina* onto which the image produced by the eye's optical system, including the crystalline lens, is projected. The space between the retina and the crystalline body is filled with a jellylike transparent *vitreous* humor. Note that the images of objects on the retina are inverted. However, thanks to the activity of the brain which receives the signals from the optic nerve, we are able to see all the objects in their natural position.

When the ciliary muscle is relaxed, the image of distant objects is focused on the retina. The construction of the human eye is such that we can see without strain only objects not nearer than six metres from the eye. The images of nearer objects are located behind the retina. To obtain a sharp image of such an object, the ciliary muscle presses the crystalline lens until the image of the object shifts to the retina and then keeps the lens in the appropriate state.

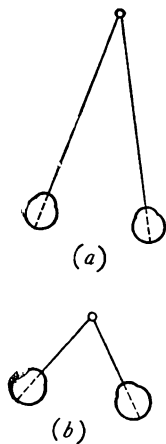
Thus, to focus the eye the ciliary muscle changes the power of the crystalline lens. Note that the lens power of the human eye is at its minimum when the ciliary muscle is relaxed, being equal to about 58 diopters. The term for the ability of the eye's optical system to form clear images of objects at different distances from the eye is *accommodation*. The rays reaching the eye from very distant objects are parallel and the eye viewing such objects is said to be accommodated to infinity. Note that this state is the least tiresome for the eye and this is why the eyes of a person deep in thought often involuntarily accommodate themselves to infinity.

The accommodation of the eye is limited. The ciliary muscle can increase the lens power of the eye by no more than 12 D. Prolonged scrutiny of close objects makes the eye tired, the ciliary muscle relaxes and the image becomes blurred.

The eyes enable a person to see clearly not only in daylight. At dusk or even at night after a man spends some time in darkness he begins to discern the shape of distant objects and to see nearby objects quite clearly. The term for the ability of the eye to adapt itself to various levels of excitation of the sensitive terminations of the optical nerve in the retina, that is, to different levels of brightness of the objects in view, is *adaptation*. The nighttime sensitivity of the human eye to luminous radiation reaching it is several billion times higher than the daytime sensitivity.

A very important faculty is a person's ability to determine with the aid of his eyesight the relative position of objects in space. The explanation for this ability is as follows. When a person looks at an object he positions the optical axes of his eyes so that they intersect at the object (Fig. 34.12). The closer the object to the man the greater the muscular strain required to make both optical axes intersect at the object. Feeling this strain the person assesses the distance to the object. The term for the process of making both optical axes intersect at an object is *convergence*. For distant objects the variation of the angle between the optical axes with the change in the object being viewed is slight, and the person can no longer correctly assess their position. When viewing very distant objects, the axes of the eyes are parallel and the person cannot even discern whether the object he is looking at is moving or at rest. Note that a definite role in assessing the position of objects is also played by the strain of the ciliary muscle, which contracts the crystalline lens when the person regards objects close to him.

Fig. 34.12 Convergence of eyes.



### 34-6 Persistence of Vision

If a glowing splinter is quickly moved in darkness (in circles, for instance), a distant observer will see a bright ring. Research into phenomena of this sort has established that a person continues to see an object on average for 0.1 s after it has been removed from his sight. This is the *persistence of vision*, which is of great practical importance.

If a man quickly runs past a fence with narrow slits between the boards, he sees everything behind the fence

clearly. The term for this effect is *stroboscopic* (from the Greek *strobos* for whirling round).

The stroboscopic effect is the basic principle of the cinema. It is also used in television to produce a sensation of motion across the screen of the cathode-ray tube. The pictures on the cinema screen change about 20 times per second. During the change-over the objective of the projector camera is closed and the screen is not illuminated. However, the audience does not notice this. It sees the changing pictures on the screen in a continuous sequence. This is how the effect of motion is reproduced on the screen.

### 34-7 Angle of View

When looking at a row of telegraph poles, we perceive distant poles as being smaller in height than the less distant ones. The explanation is that although the height of all the poles is identical, the size of the image of a nearby pole on the retina is greater than that of a distant one. Actually the size of the image of an object on an eye's retina is fully determined by the angle of sight,  $\varphi$  (Fig. 34.13).

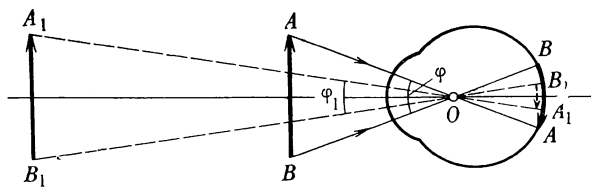
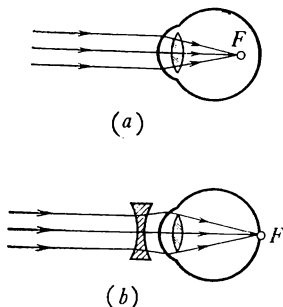


Fig. 34.13 The farther the object from the eye, the smaller the angle of view ( $\varphi_1 < \varphi$ ).

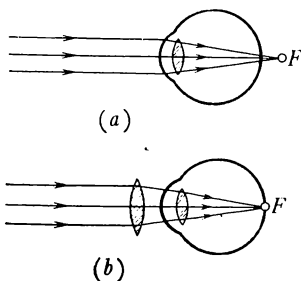
The term *angle of view* applies to the angle between straight lines drawn from the optical centre of the eye to extreme points of the object. When this angle is large, the image covers a large number of terminations of the optical nerve in the retina (light-sensitive cells) and the person discerns numerous details of the image in view. Obviously, the size of the image (of the angle of view) depends on the size of the object and on the distance to it.

As the distance of the object from the eye increases its image on the retina decreases in size until the whole image occupies the area of no more than one light-sensitive cell and man perceives it as a bright spot. This corresponds to an angle of view of approximately  $30''$ . However, in practice a person is not able to discern details of an object when the angle of view is less than one minute. Therefore in calculations one should assume the maximum angle of view for

**Fig. 34.14** (a) In nearsighted eye, image is formed before retina; (b) diverging lens corrects this defect of eyesight.



**Fig. 34.15** (a) In farsighted eye, image is formed behind retina; (b) converging lens corrects this defect of eyesight.



which the objects are seen as a single object to be equal to one minute. It is sometimes termed the *minimum angle of view*.

For viewing distant or very small nearby objects optical instruments are used, which greatly increase the angle of view.

### 34-8 Defects of Vision. Optical Illusions

It was established in the preceding section that to increase the angle of view one must bring the object closer to the eye. However, with the object too close the eye soon becomes tired, a normal eye being generally unable to see objects clearly at distances below 20 cm.

The minimum distance at which the eye is able to see objects clearly without excessive fatigue is termed distance of *maximum visual acuity* ( $L$ ). The accepted value for people with normal eyesight is 25 cm. This is the distance at which people hold books in front of their eyes.

Observations show that for some people the distance of maximum visual acuity is below 25 cm. Such people are said to be *nearsighted*. For others the optimum distance is greater than 25 cm. They are said to be *farsighted*. Note that nearsighted people are unable to see distant objects clearly and farsighted people nearby objects. Therefore, while scrutinizing an object a nearsighted person brings it closer to his eyes and the farsighted one, on the contrary, moves it away.

Persons with such defects use eyeglasses. The principal focus  $F$  of the optical system of a nearsighted eye is in front of the retina (Fig. 34.14a). Eyeglasses with diverging lenses selected so as to shift the principal focus of the system to the retina are the remedy for such eyes (Fig. 34.14b). The principal focus  $F$  in the eyes of a farsighted person lies behind the retina (Fig. 34.15a). The remedy for them are eyeglasses with converging lenses (Fig. 34.15b).

Theory shows that the glasses shift the position of the system's principal focus without changing the lens power of the system as a whole. This may be proved with the aid of formula (34.3) if one takes into account that each lens of the glasses is located at the front principal focus of the eye's system (substituting  $a = f_2 = 1/P_2$  into (34.3), we obtain  $P_s = P_2$ ).

The process of vision is a very intricate one. A person's perception of visual sensations depends on the activity of the brain. Visual images are formed in the brain and are connected with psychic processes. For this reason identical

objects do not always produce an identical sensation. In some cases these variations are due to the environment, for instance the contrast of illumination, the colour of surrounding bodies, their position in space, etc. There may be cases

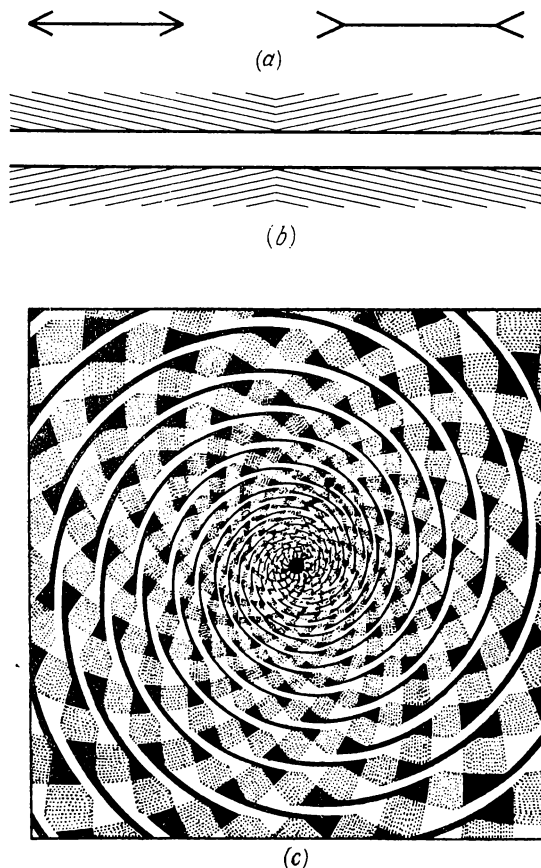


Fig. 34.16 Optical illusions: (a) segments appear to be of different length; (b) lines look as if not parallel; (c) concentric circles look like spirals.

when a person wrongly identifies objects, that is, mistakes one object for another. Such erroneous perception is termed *optical illusion*. There are many kinds of optical illusions, which sometimes may be the cause of an error. Let us cite some examples.

Figure 34.16a shows two segments of equal length which because of the arrows at their ends appear to be of different length. Figure 34.16b depicts parallel straight lines, but oblique shading creates the illusion that they are converging. The pattern shown in Fig. 34.16c is made up of circles, but



the variegated background creates the impression that they are spirals. These examples prove that a person should not always believe his visual impressions.

### 34-9 The Magnifying Glass

The term *magnification of an optical instrument* applies to the ratio of the angle  $\varphi$  at which the eye sees the image of an object in an instrument to the angle of view  $\varphi_0$  at which the eye sees the same object without the instrument from the distance of maximum visual acuity (for close range instruments) or from the same distance as the instrument (for long-range instruments):

$$N = \frac{\varphi}{\varphi_0} \quad (34.8)$$

Since angles  $\varphi$  and  $\varphi_0$  are usually small, the magnification of an optical instrument is often found with the aid of the approximate formula

$$N \approx \frac{\tan \varphi}{\tan \varphi_0} \quad (34.9)$$

One of the simplest types of optical instruments is the magnifying glass — a converging lens used to obtain magnified images of small objects.

The object viewed with the aid of a magnifying glass is usually placed in the focal plane of the lens or a little nearer. Figure 34.17*a* depicts a small object  $AB$  and its image in the eye  $A_1B_1$ . If the object  $AB$  is at the distance of maximum visual acuity,  $L$ , from the eye, its angle of view is  $\varphi_0$ . We now place a magnifying glass in front of the eye and move the object  $AB$  to the focal plane of the lens (Fig. 34.17*b*). In this situation a parallel beam of rays from every point of the object  $AB$  reaches the eye. The eye's optical system projects them onto the retina where the image  $A_2B_2$  is obtained. Since in this case the object  $AB$  is viewed at an angle  $\varphi$  exceeding  $\varphi_0$ , the image  $A_2B_2$  is larger than  $A_1B_1$ . A person can then resolve details of the object  $AB$  which he was unable to see with the naked eye. The magnification of the magnifying glass will in this case be expressed by the formula

$$N = \frac{\tan \varphi}{\tan \varphi_0} = \frac{AB}{f} \div \frac{AB}{L} = \frac{L}{f}$$

Since for people with normal eyesight  $L = 0.25$  m, we finally obtain the formula for the magnification of a magni-

ying glass:

$$N = \frac{0.25}{f} \quad (34.10)$$

Note that in the case described above a person's eye is accommodated to infinity and this enables him (or her) to view the object  $AB$  through a magnifying glass without effort and for a long time without experiencing fatigue.

By moving the object  $AB$  from the focal plane closer to the lens (Fig. 34.17c) one can obtain its virtual image  $A'B'$  at

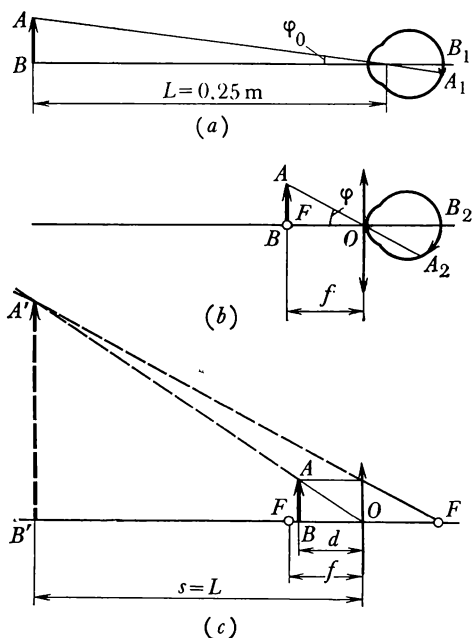


Fig. 34.17 (a) At distance of maximum visual acuity small object is seen without lens at angle  $\varphi_0$ ; (b) object is in focal plane of lens; (c) object is between lens and its focal plane.

the distance of maximum visual acuity,  $L$ . Let us find the magnification of the magnifier in this case. Since in the latter case the angle of view will be greater than in the former, the magnification will also be greater. It follows from Fig. 34.17 that  $\tan \varphi = AB/d$

$$N = \frac{\tan \varphi}{\tan \varphi_0} = \frac{AB}{d} \div \frac{AB}{L} = \frac{L}{d} = \frac{1}{d} L$$

Since  $L$  is always positive and  $s$  in this case is negative, it follows that  $L = -s$  and we obtain from (33.3):  $1/d - 1/L = 1/f$ , whence  $1/d = (L + f)/Lf$ . Hence

$$N = \frac{(L + f)L}{Lf} = \frac{L + f}{f}, \quad \text{or} \quad N = \frac{L}{f} + 1 \quad (34.11)$$

Thus, if the eye is accommodated to the distance of maximum visual acuity, the magnification of the magnifying glass is greater by unity than when it is accommodated to infinity. However, in the former case the eye is strained and soon becomes tired. Accordingly, when a person looks through a magnifier for an appreciable time, the eye soon accommodates itself to infinity. This means that for practical purposes the magnification should be determined from formula (34.10).

### f34-10 The Microscope

The instrument which enables large magnifications of small objects to be obtained is called a *microscope* (Fig. 34.18a). It consists of two high-power lenses. The lens in front of which the object is placed is termed the *objective* ( $O_1$ ) and the one viewed by the eye is termed the *eyepiece* ( $O_2$ ).

The objective and the eyepiece are self-contained optical systems enclosed in separate mountings inserted into a metal tube (T). The object being viewed is placed on the specimen table S and illuminated from below with the aid of a mirror M and a lens system L. To obtain a sharp image the tube is moved with the aid of adjustment screws  $A_1$  and  $A_2$ .

The path of rays in the microscope is shown in Fig. 34.18b. The object  $AB$  is placed just a little behind the principal focus of the objective. The eyepiece is positioned to have the inverted real image  $A_1B_1$  of the object in its principal focus and acts as a magnifying glass. We now examine factors determining the magnification of the microscope.

A person looking into the eyepiece sees the image  $A_1B_1$  at an angle  $\varphi$ . It can be seen from the diagram (Fig. 34.18b) that

$$\tan \varphi = \frac{A_1B_1}{f_{ep}} \quad \text{and} \quad A_1B_1 = (a - f_{ep}) \tan \varphi_1$$

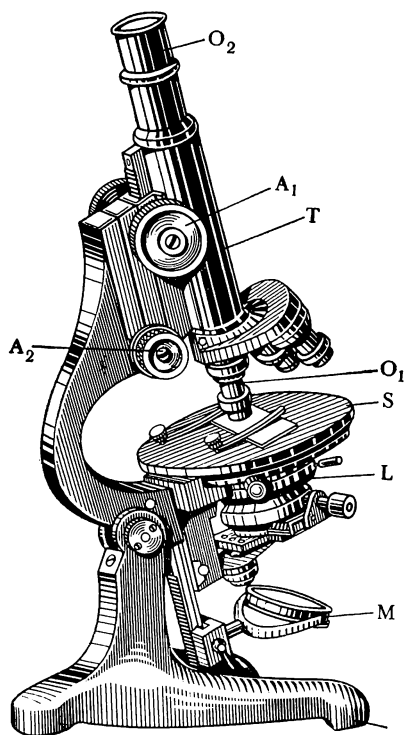
Since  $\tan \varphi_1 \approx \frac{AB}{f_{ob}}$ , it follows that  $A_1B_1 = \frac{(a - f_{ep}) AB}{f_{ob}}$ .  
Hence

$$\tan \varphi = \frac{A_1B_1}{f_{ep}} = \frac{(a - f_{ep}) AB}{f_{ep} f_{ob}}$$

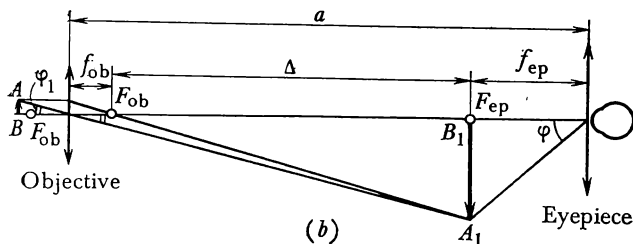
Since  $\tan \varphi_0 = \frac{AB}{L}$  (where  $L$  is the distance of maximum visual acuity), the microscope's magnification is

$$N = \frac{\tan \varphi}{\tan \varphi_0} = \frac{(a - f_{ep}) AB \times L}{f_{ep} f_{ob} \times AB} = \frac{(a - f_{ep}) L}{f_{ep} f_{ob}}$$

Taking into account the fact that the focal length of the objective is quite small, one can assume the quantity  $(a - f_{ep})$  to be approximately equal to the distance between



(a)



(b)

the foci of the objective and the eyepiece, normally denoted  $\Delta$  and termed the *length of the microscope tube*. The formula for the microscope's magnification then becomes

$$N = \frac{0.25}{f_{ep}} \frac{\Delta}{f_{ob}} \quad (34.12)$$

**Fig. 34.18** (a) Microscope; (b) path of rays in microscope.

Since  $N_{ep} = 0.25/f_{ep}$  is the magnification of the eyepiece and  $N_{ob} = \Delta/f_{ob}$  is the magnification of the objective, it can be said that the magnification of a microscope is the product of magnifications of its objective and eyepiece.

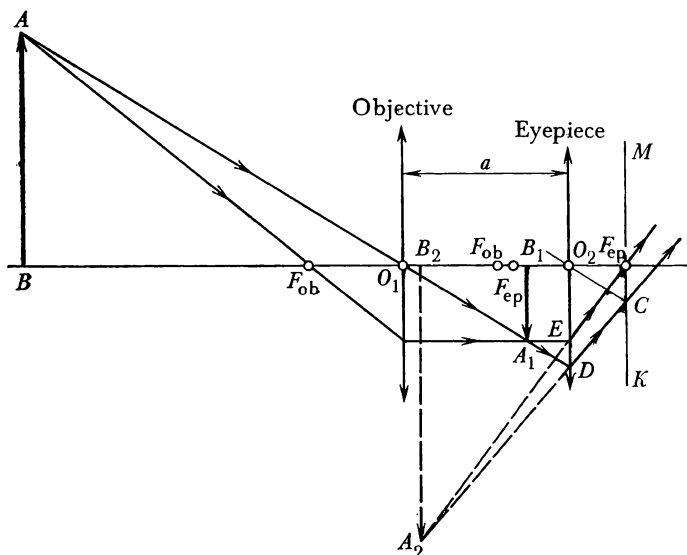
$$N = N_{ob}N_{ep} \quad (34.13)$$

Note that a person sees in a microscope a virtual inverted and magnified image of the object. Optical microscopes have magnifications not exceeding 1000.

### 34-11 Telescopes

The term for an optical instrument intended for viewing distant objects which cannot be brought closer to the eye is the *astronomical telescope*. The first telescopes were built in 1609 by Galileo Galilei and by the German astronomer Johannes Kepler (1571-1630).

Fig. 34.19 Path of rays in Kepler telescope.



The telescope in which an increase in the angle of view is obtained with the aid of lenses is termed the *refracting telescope*. A telescope in which a similar effect is obtained with the aid of mirrors is termed the *reflecting telescope*.

The *Kepler telescope* consists of two converging lenses: an objective and an eyepiece. The path of rays in the Kepler telescope is depicted in Fig. 34.19. Note that  $O_2C$  is parallel to  $AA_1$  and that  $MK$  is the focal plane of the objective. The

dimensions of the objective are usually large and its lens power small, the eyepiece acting like a magnifying glass through which the image formed by the objective is viewed. The image of the object is located at the principal focus of the objective,  $F_{ob}$ , while the eyepiece is positioned to have

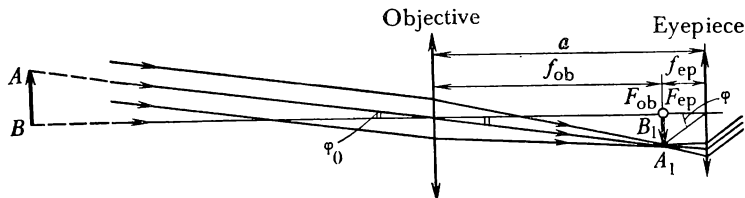


Fig. 34.20 Magnification of Kepler telescope.

this image in its principal focus,  $F_{ep}$ . Accordingly, the distance  $a$  between the objective and the eyepiece in this instrument is equal to the sum of focal lengths  $f_{ob}$  and  $f_{ep}$ , that is, the length of the Kepler telescope is

$$a = f_{ob} + f_{ep} \quad (34.14)$$

Let us find the magnification of the Kepler telescope. Figure 34.20 shows rays coming from a very distant object seen by the naked eye at an angle  $\varphi_0$ . When a person looks at this object through a Kepler telescope, he (or she) sees its image  $A_1B_1$  at an angle  $\varphi$ . Since  $\tan \varphi = A_1B_1/f_{ep}$  and  $\tan \varphi_0 = A_1B_1/f_{ob}$ , it follows that the magnification of the Kepler telescope is

$$N = \frac{\tan \varphi}{\tan \varphi_0} = \frac{A_1B_1}{f_{ep}} \div \frac{A_1B_1}{f_{ob}}, \quad \text{or} \quad N = \frac{f_{ob}}{f_{ep}} \quad (34.15)$$

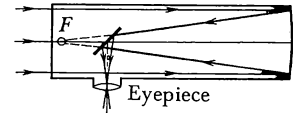
From (34.15) it follows that in order to obtain large magnifications the objective in the Kepler telescope must be a long-focus one and the eyepiece a short-focus one. (Note that one should not identify the magnification of an optical instrument,  $N$ , with lateral magnification,  $\beta$  (see Section 33-8).)

The diameters of objectives in modern refracting telescopes exceed one meter, their focal length being almost 20 m.

Figure 34.21 shows a schematic diagram of a reflecting telescope with its focus at point  $F$ . The eyepiece is on the side. Light rays reach the eyepiece after being reflected by a plane mirror. The aperture of the telescope's mirror may be as great as 5 m.

The telescope not only helps to resolve objects at small angular distances from one another but also makes it possible to observe very weak light sources, since the objective collects a wide beam of rays (immeasurably wider than that collected by the pupil of the eye).

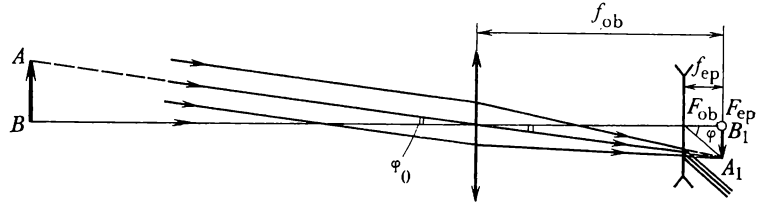
Fig. 34.21 Path of rays in reflecting telescope.



### 34-12 Galileo's Telescope and Binoculars

The image of an object in the Kepler telescope is an inverted one. This is of no importance if one is observing celestial bodies, but is inconvenient for observing objects on the Earth. For this reason a lens is added, whose only purpose is

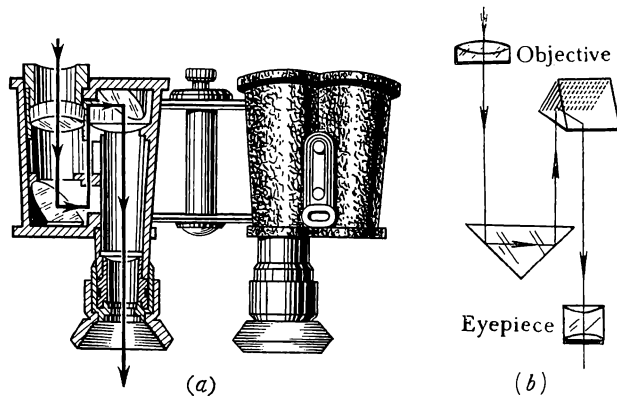
Fig. 34.22 Path of rays in Galileo's telescope.



to invert the image. This makes the telescope  $4f$  times longer, where  $f$  is the focal length of the additional lens. (Explain why  $4f$ .)

Such a telescope is too unwieldy, and so in practice *Galileo's telescope* is used instead. It consists of a converging lens (objective) and a diverging lens (eyepiece). A diagram

Fig. 34.23 Prismatic binoculars.



of Galileo's telescope is presented in Fig. 34.22. In it the positions of the objective and the eyepiece are such that their focal points coincide, for reasons discussed in Section 34-11. Formula (34.14) remains valid with allowance made for the negative sign in front of  $f_{ep}$ . The magnification of Galileo's telescope is found from formula (34.15) and turns out to be less than that of the Kepler telescope. Two Galileo's telescopes joined together form opera glasses, handy because of their small size.

Prismatic binoculars (Fig. 34.23a)—a combination of two Kepler telescopes—have a greater magnification. The image in these binoculars is inverted not with the aid of a lens but with the aid of two total reflection prisms (Fig. 34.23b) in each of the telescopes. This makes it possible to combine the comparatively small dimensions of prismatic binoculars with a substantial magnification.

## Phenomena Arising from Wave Nature of Light

## 35

### 35-1 Interference of Light

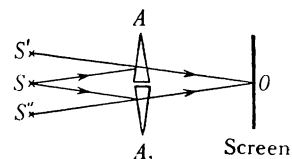
This chapter deals with those phenomena which the corpuscular theory of light is unable to explain. They include interference, diffraction and polarization of light. It was these phenomena the study of which helped Fresnel establish the wave nature of light and to prove that luminous radiation consists of transverse waves.

Since only waves from coherent sources experience interference, Fresnel began by devising methods for producing coherent light sources. It was established by experiment that the radiation of two individual light sources, even if one is an exact copy of the other, does not experience interference. This means that such sources are incoherent. Different beams of coherent radiation could be produced by partitioning the radiation from a single light source.

To obtain an interference pattern the rays emanating from some light source in different directions should be made to overlap with the aid of some optical device. To this end Fresnel used mirrors and prisms. Figure 35.1 shows the schematic diagram of the Fresnel biprism used to obtain two coherent light sources.

The bases of two identical glass prisms,  $A$  and  $A_1$ , with very small angles are glued together. If a source of light  $S$  is placed on one side of the prism and a screen on the other, the interference pattern can be observed on the screen. The explanation is that all rays falling on prism  $A$  after refraction follow paths they would have followed if they had originated at point  $S'$ , which is the virtual image of the light source  $S$ . Similarly, the rays falling on the prism  $A_1$  and refracted in it follow paths they would have followed if they had originated at  $S''$ . The rays, appearing to emanate

Fig. 35.1 Fresnel biprism; coherent rays from virtual sources  $S'$  and  $S''$  interfere on screen.



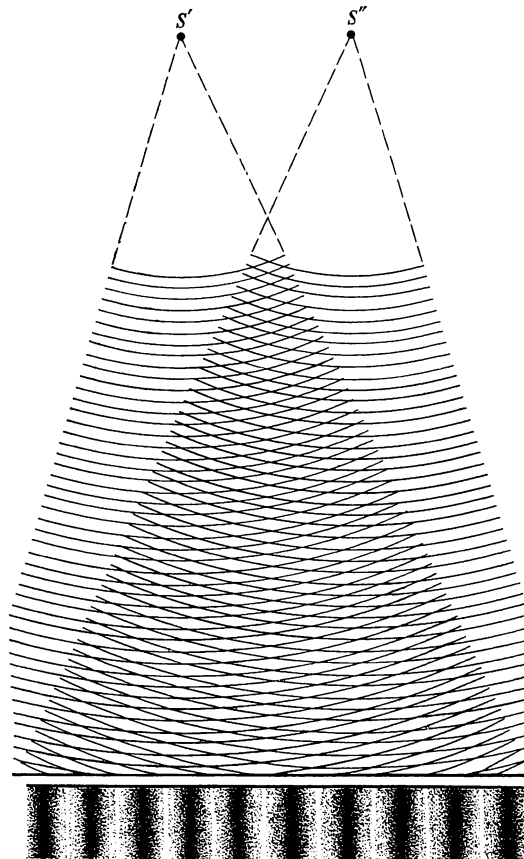


from two virtual coherent light sources,  $S'$  and  $S''$ , overlap over the whole surface of the screen (Fig. 35.2).

The interference pattern on the screen will be most distinct if the light source  $S$  emits *monochromatic radiation*, that is, radiation of a definite frequency. Such radiation can be obtained with the aid of lasers (see Section 38-18). Radiation in a comparatively narrow frequency band can be obtained with the aid of colour filters—special glasses transmitting light of only one colour.

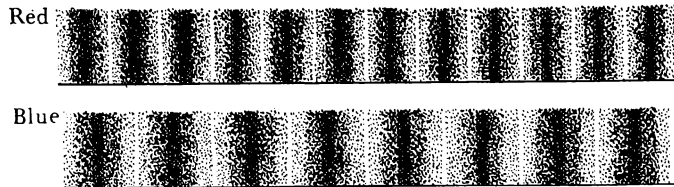
If the light source  $S$  is made in the form of a narrow bright slit normal to the plane of Fig. 35.1, alternating dark and bright stripes termed interference fringes will be seen on the screen. The fringe at point  $O$  of the screen opposite the light source  $S$  will be bright because at this point the coherent rays overlap in-phase. (Why?) Along both sides away from the central bright fringe  $O$  on the screen the difference

**Fig. 35.2** Interference of monochromatic rays coming from luminous slits  $S'$  and  $S''$  (interference pattern seen on screen is shown below).



in wave paths increases, and when it reaches  $\lambda/2$  the stripes turn dark (see Section 27-21). When the difference in wave paths reaches  $\lambda$ , two bright fringes appear on the screen. And so it continues. Thus, the interference pattern on the screen is made up of alternating bright and dark fringes spaced at approximately equal distances.

It can easily be seen that with the whole setup remaining stationary the distance between two adjacent bright (or dark) fringes would depend on the wavelength  $\lambda$ : the smaller  $\lambda$



**Fig. 35.3** Separation of interference fringes depends on wavelength of monochromatic rays; interference pattern proves that wavelength of red rays is greater than that of blue.

is the smaller the distance across the screen along which the difference in ray paths changes by a whole wavelength, that is, the denser the interference fringes on the screen. For instance, when the biprism is illuminated with red light, the distances between the fringes are greater than when it is illuminated with blue light (Fig. 35.3). Note that dense shading in Fig. 35.3 denotes a bright fringe and sparse shading a dark fringe. Point  $O$  marks the central bright fringe, the path difference for which is zero.

Such experiments show that there is a definite colour corresponding to a definite frequency band, that is, that colour is determined by the frequency of luminous radiation. The colours of monochromatic light change with the increase in wavelength in the following order: violet, blue, green, yellow, orange and red.

If the biprism is illuminated with white light, there will be a white fringe at point  $O$  (see Fig. 35.3) and coloured fringes of all the colours present of the rainbow on both sides of it. This experiment demonstrates that white colour is a composite colour, that is, it is made up of a mixture of rays of all wavelengths of visible light.

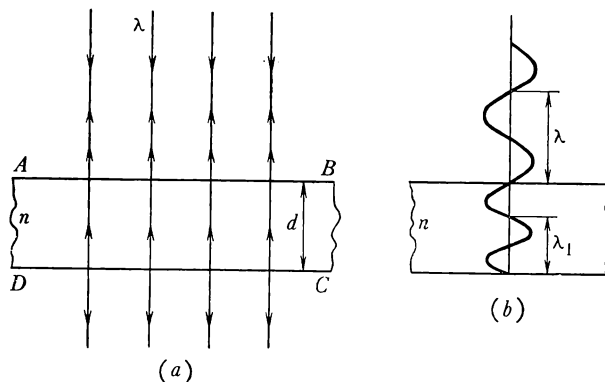
### 35-2 Colours of Thin Films

Everyone of us used to blow soap bubbles with continuously changing colours. This phenomenon is due to the interference of light in thin transparent films whose thickness does not exceed several micrometres.

Let us first examine interference in a plane parallel plate. (The term plane parallel applies to a plate whose plane surfaces are parallel.) Good quality window glass or mirror glass serves as examples of such a plate.

Let a parallel beam of monochromatic rays fall on a very thin plane parallel plate at right angles to it (Fig. 35.4a). The light rays are partially reflected from surface  $AB$ , some of them penetrating into the plate. The process is

Fig. 35.4. Diagram explaining interference in plane parallel plate.



repeated at the surface  $CD$ . Since the ray reflected from the surface  $CD$  on emerging from the plate takes the same path as the ray reflected from the surface  $AB$ , both rays are coherent and therefore experience interference.

Note that in the case just discussed the conditions for interference are identical for all rays striking the surface of the plate at any of its points. Therefore, if the interfering rays overlap out-of-phase, the whole plate will appear dark, if they overlap in-phase the whole plate will be illuminated by light of the colour corresponding to the wavelength of the monochromatic rays  $\lambda$ .

The interference of the rays depends on their optical path difference. This is not the same as geometrical path difference. Consider the case when the interference of reflected light is observed, that is, when the observer looks at the plate from above (Fig. 35.4). The geometrical path difference will be  $2d$ , since the ray reflected from the lower face of the plate, having to first go down and then up, covers additional distance equal to the double thickness of the plate. However, the rays have the wavelength  $\lambda$  in air, their wavelength in glass changing in proportion to the change in light velocity in the plate as compared with air:

$$\frac{c}{f} = \frac{\lambda f}{\lambda_1 f} = \frac{\lambda}{\lambda_1}$$

where  $v$  and  $\lambda_1$  are the light propagation velocity and the wavelength in the plate's material, and  $f$  is frequency. Since  $c/v = n$ , it follows that  $\lambda/\lambda_1 = n$  and

$$\lambda_1 = \lambda/n \quad (35.1)$$

Since  $n$  exceeds unity, the wavelength is shorter in the plate (see Fig. 35.4*b*). Therefore the path difference of the interfering rays will not be  $2d$  but  $2dn$ . Moreover, in optics, as in mechanics (see Section 27-19), reflection from a medium of greater density (in our case, optical density) is accompanied by a loss of half a wave, the reflection from a medium of lower density not being accompanied by such a loss. In the case being considered the half-wave is lost in reflection from the upper face. Hence, the optical path difference will in our case be

$$\Delta = 2dn - \lambda/2$$

We recall that the maximum intensification takes place when the wave path difference is equal to an even number of half-waves. Hence, the condition for the maximum intensification of interfering rays reflected from a plate is expressed by the relation

$$\Delta = 2dn - \frac{\lambda}{2} = 2k \frac{\lambda}{2}, \quad \text{or} \quad 2dn = (2k+1) \frac{\lambda}{2} \quad (35.2)$$

where  $k$  is an integer (1, 2, 3, . . .).

It can easily be seen that the condition for the maximum suppression of light intensity is expressed by the relation

$$\Delta = 2dn - \frac{\lambda}{2} = (2k-1) \frac{\lambda}{2}, \quad \text{or} \quad 2dn = 2k \frac{\lambda}{2} = k\lambda \quad (35.3)$$

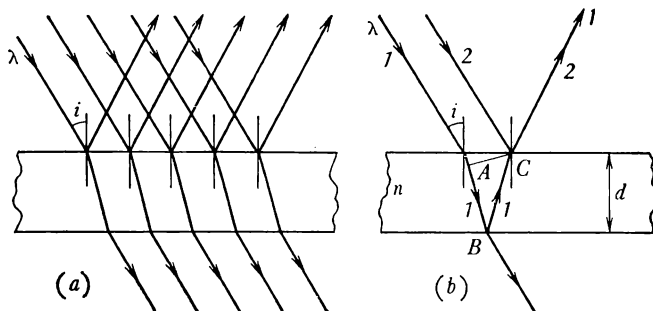
If the plate is seen in transmitted light, that is, from below, the conditions are reversed: the relation (35.3) expresses the condition for maximum intensification and the relation (35.2) for maximum suppression of light.

A change in the angle of incidence of the rays on the plate to an angle  $i$  (Fig. 35.5*a*) changes the path difference of the interfering rays. For the rays 1 and 2 it will be equal to  $(AB + BC)n - \lambda/2$  (Fig. 35.5*b*). It should be noted here that  $AC$  is the position of the wavefront at the moment the ray 2 is reflected from point  $C$  ( $AC \perp AB$ ). The optical path difference is shown to decrease with the increase in the angle of incidence,  $i$ . This means that as the plate is inclined to the rays, it will appear dark and bright in succession.

If the plate is illuminated with white light, the interference of rays of one wavelength will result in their intensi-

fication and of another in suppression. Accordingly, the plate will appear to the observer to be coloured, the colour being that of the rays with maximum intensification.

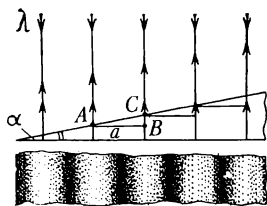
**Fig. 35.5** Interference also takes place when parallel rays fall on plate at oblique angle: (a) both reflected and refracted rays can interfere; (b) interference of reflected rays 1 and 2.



Obviously the colour of the plate will change as it is inclined to the rays. It should be stressed that the above refers to the case where the rays falling on the plate are parallel.

### 35-3 Interference in a Wedge-Shaped Film. Newton's Rings

**Fig. 35.6** Interference of monochromatic rays reflected from wedge-shaped film results in alternating bright and dark fringes.



Let us look into the details of the interference of light in a wedge-shaped film with a very small angle of the wedge; the refractive index is  $n$ . When such a film is illuminated with monochromatic light falling at right angles to one of the wedge's faces, alternating dark and bright fringes parallel to the edge of the wedge are seen on its surface (Fig. 35.6). Let us find out how these fringes come about.

Figure 35.6 depicts the path of interfering rays in the wedge. It can be seen that as the distance from the wedge's edge increases, the path difference of the rays increases as well. Let there be maximum intensification of light at point A. In that case there will be a point B at some distance  $a$  from point A where, because of an increase in the thickness of the wedge, there will again be maximum intensification of light. Since the increase in the path difference should in this case be  $\lambda_1$ , it follows that  $2BC = \lambda_1$ . Since  $\lambda_1 = \lambda/n$ , we obtain

$$2BC = \lambda/n$$

It may be seen from  $\triangle ABC$  that  $BC = a \tan \alpha$ ; therefore

$$2a \tan \alpha = \lambda/n$$

It is an established principle of trigonometry that for small angles the tangent of an angle can be approximated by the angle itself expressed in radians; therefore

$$2a \alpha = \lambda/n$$

whence

$$a = \lambda/2n\alpha \quad (35.4)$$

It can easily be seen that the next bright fringe will be a distance  $a$  away from point  $C$ , and so on. This means that in this case the interference fringes are equidistant from one another.

It follows from relation (35.4) that an increase in the angle  $\alpha$  results in a decrease in the spacing between the bright (and the dark) fringes. If the angle  $\alpha$  is gradually decreased, the interference fringes will separate, vanishing altogether as the faces of the film become parallel. On the contrary, when the angle  $\alpha$  is increased, the fringes draw closer and at an angle of about  $1^\circ$  they overlap, that is, the interference pattern vanishes.

When a wedge-shaped film is illuminated with white light, fringes appear coloured with all the colours of the rainbow. A similar phenomenon can be observed when a soap bubble is illuminated with white light. The variation in the colour of the soap bubble is due to the variation in the thickness of the film, caused by water streaming down into the lower part of the bubble.

A convenient device for observing the interference of light is a plano-convex lens placed on a plane parallel plate so that a wedge-shaped air gap is formed between the lens and the plate (Fig. 35.7a).

For the interference to be clearly visible the radius of curvature of the convex lens surface must be large enough. If this device is illuminated with a parallel beam of monochromatic light falling at a normal to the plane surface of the lens (Fig. 35.7a), alternating dark and bright interference rings termed *Newton's rings* will be distinctly seen in reflected light (Fig. 35.8). In this case the interfering rays are those reflected from the curved surface of the lens and from the face of the plate. For instance, the ray 1 (Fig. 35.7a) is the result of interference of rays reflected from points  $A$  and  $B$ .

Since the locus of equal air-gap thickness, say  $AB$ , is the circumference of radius  $OB$ , the interference pattern will be in the form of rings. The rings draw closer together with the increase in their radius, since in this direction the angle of the air wedge increases (Fig. 35.7b).

Fig. 35.7 Diagram explaining Newton's rings: (a) the conditions for interference of rays are identical on circumference with radius  $OB$ ; (b) wedge-shaped gap of increasing angle corresponds to interference rings of greater radius.

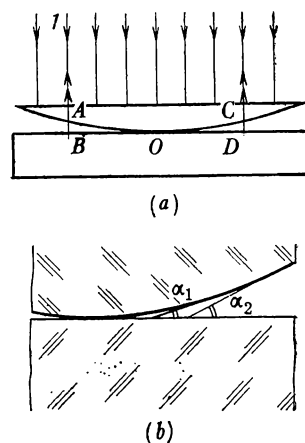
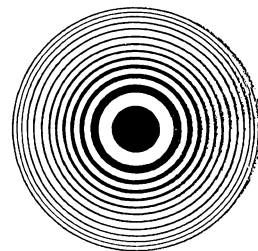


Fig. 35.8 Photograph of Newton's rings.

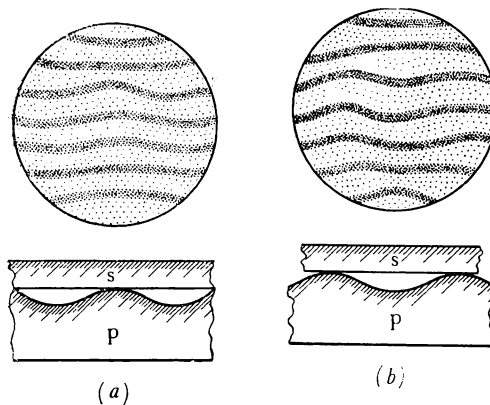


Note that the illumination of this device with white light results in the appearance of rings coloured in all colours of the rainbow. (What difference, apart from colour, will there be between the interference patterns produced first by red and then by blue light?)

### 35-4 Interference in Nature and Technology

In natural conditions light interference phenomena are responsible for the colouring of thin films of oil spread over the surface of water or asphalt, for the colouring of wings of some insects (for instance, dragon flies and butterflies), etc. Note that the difference in colour of the films is due to the differences in their thickness.

Fig. 35.9 Checking quality of surface by interference method: (a) curved interference fringes showing convex spot on surface p; (b) curved interference fringes showing concave spot on surface p.



In modern science and technology interference is widely used for making precise measurements and for estimating the quality of a surface finish, this being especially important in manufacturing optical glasses for instruments and in many other applications. The interference method of measuring the wavelength of luminous radiation guarantees accuracy up to eight significant digits. This method was used to measure the length of the standard metre. The result was a new definition of the metre: the *metre* is a length which contains 1650 763.73 wavelengths (in a vacuum) of orange rays radiated by krypton vapour.

When the quality of the surface polish is of critical importance, the following method is used (Fig. 35.9). A standard

plate *s* is placed on top of the surface *p* being checked, and the plate is illuminated with monochromatic light. A high quality of the surface produces parallel interference fringes. Any roughness curves the interference fringes (Fig. 35.9).

Interference makes it possible to measure the thickness of very thin films, very thin threads and very small angles. Those are by no means all the applications of interference in modern technology.

### 35-5 Diffraction of Light

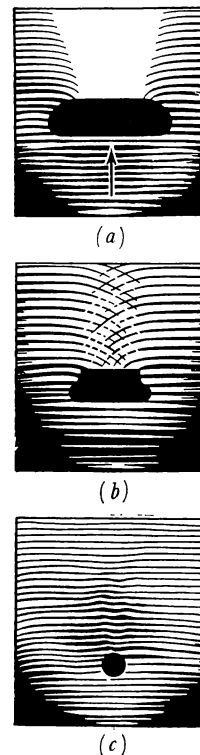
The second proof of the wave nature of light is the phenomenon of *diffraction* (from the Latin *diffractus* for break to pieces). The term diffraction applies to waves bending round obstacles. The diffraction of waves travelling along a water surface is illustrated in Fig. 35.10. When the obstacle is large (compared with the wavelength), there are no waves behind it (Fig. 35.10*a*). When the obstacle is small, the waves bend round its edges (Fig. 35.10*b*). When the size of the obstacle is very small, the waves bend round it so that it has practically no effect on the wavefront (Fig. 35.10*c*).

Figure 35.11 illustrates the passage of waves through a hole in the obstacle. When the hole is large (compared with the wavelength) the waves hardly bend round its edges (Fig. 35.11*a*). When the hole is small, the bending effect round the edges is noticeable (Fig. 35.11*b*). When the hole is very small, the waves spread all over the surface behind the obstacle (Fig. 35.11*c*). In this case the hole appears to act as an independent source of waves, which propagate in all directions behind the obstacle.

The explanation for all these phenomena is that the obstacle cuts off a part of the wavefront. The size of the obstacle or of the hole for which diffraction becomes noticeable depends on the wavelength: the smaller the obstacle in comparison with the wavelength the more manifest is the diffraction.

When the size of the obstacle (or hole) is comparable to the wavelength, the diffraction is observed quite close to the obstacle (Figs. 35.10*c* and 35.11*c*). However, when the obstacle is large compared to the wavelength, diffraction still can be detected but at greater distances from the obstacle. The explanation is that the variations in the wavefront introduced by the obstacle become more noticeable as the distance from the obstacle increases. Thus, the larger the obstacle the greater the distance at which the diffraction

Fig. 35.10 Diffraction of waves on surface of water: (a) wave propagating near large obstacle; (b) waves bend around edges of smaller obstacle; (c) waves bend around small obstacle.



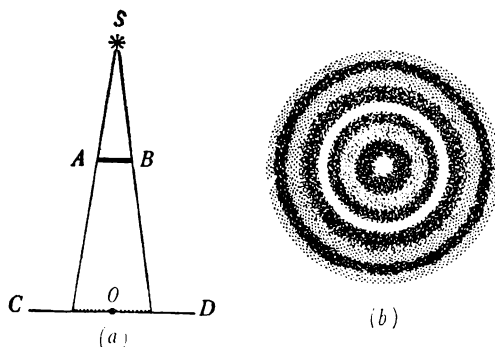


phenomenon is detected. However, the energy of the waves should be great enough for diffraction to be felt.

Let us consider now the diffraction of light. Since the wavelengths of luminous radiation are very short, the diffraction of light can be observed only at substantial distances from an obstacle or a hole. Let a very small disk of diameter  $AB$  stand in the path of rays emanating from a point light source  $S$  (Fig. 35.12*a*), diffraction being observed on the screen  $CD$ . Should the light propagate along straight lines, there would be a shadow of diameter  $CD$  on the screen. Yet when the distance from the disk to the screen is great enough, a diffraction pattern is obtained on the screen in the form of alternating dark and bright rings (Fig. 35.12*b*), there being a bright circular spot at the centre of the screen, that is, at point  $O$ .

Calculations show that the energy of oscillations from only a part of the wave surface directly adjoining the disk  $AB$  arrives at point  $O$ . All other oscillations originating from other parts of the wavefront are suppressed as a result of interference. It may be seen in Fig. 35.12*a* that all points of the wavefront encircling the disk  $AB$  are equidistant from point  $O$ . This means that the oscillations excited by them at point  $O$  should be in-phase, that is, should augment each other. Therefore there should be a bright spot at point  $O$ .

The following experiment makes it possible to observe diffraction from a narrow slit. A nontransparent screen with a narrow slit is placed in the path of parallel monochromatic rays and the diffraction is observed on another screen placed at some distance from it (Fig. 35.13). In this case a bright fringe is seen opposite the slit. Its width increases with the decrease in the width of the slit (why?), there being alternating dark and bright fringes on both sides of the central bright fringe.



Note that when obstacles and holes are illuminated with white light the diffraction pattern obtained is not so sharp and has the colours of a rainbow.

### 35-6 The Diffraction Grating. Measurement of Wavelength

In practice the observation of diffraction from a single slit (from a single hole) is difficult because very little light passes through a narrow slit. To make the diffraction pattern bright enough, light should be made to pass through several parallel slits (Fig. 35.14). In this case the diffraction phenomenon will be augmented by the interference phenomenon because the rays of a point source coming through different slits are coherent. (Why?) Maximum intensification of the brightness of monochromatic light will, obviously, be in those parts of the screen where the light arrives in-phase from all the slits.

Thus, when the illuminated slits are great in number, bright narrow lines on a dark background are visible on the screen. It has been established that the greater the number of slits and the closer they are spaced the brighter and the narrower are the places on the screen in which the rays overlap in-phase. Note that bringing the slits closer together results in an increase in the distance between the bright lines on the screen. The phenomena described above form the basis for the design of diffraction gratings.

The term for a great number of parallel and very closely spaced slits is *diffraction grating*. The gratings are fabricated either of a transparent solid material or of a good reflecting metal. In both cases parallel rulings are inscribed on the surface with the aid of a diamond cutter. The rulings made by the cutter have a rough surface, which scatters the rays. The intervals between the rulings remain transparent, or reflecting, that is, they play the part of slits. Gratings cut on a mirror surface are sometimes termed *reflecting gratings*. Modern gratings have more than a thousand rulings per millimetre, the total number being as high as a hundred thousand.

The important feature of a grating is its spacing,  $d$ , the distance from the edge on one side of a slit to the edge on the same side of the adjacent slit (Fig. 35.15*a*). Let a bundle of parallel monochromatic rays fall on a grating at a normal to its surface (Fig. 35.15*b*). In that case, as a result of diffraction, the waves on the opposite side of the grating propagate in all directions. Interference results in the intensi-

Fig. 35.13. Diffraction from single slit.



Fig. 35.14 (a) Diffraction of monochromatic light from five slits; (b) diffraction from twenty slits.

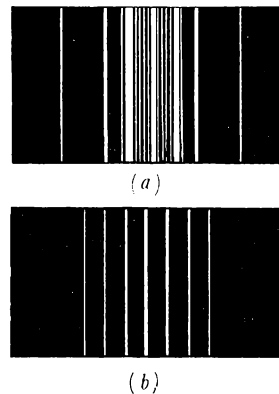


Fig. 35.15 (a) Lateral cross section of diffraction grating greatly magnified; (b) when diffraction grating is illuminated with monochromatic light, bright fringe appears at point  $M$  on screen.

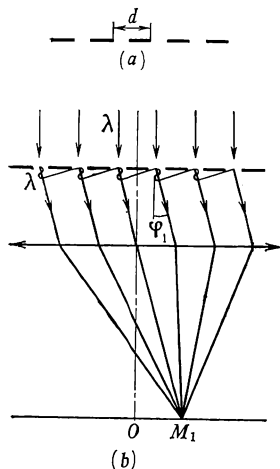
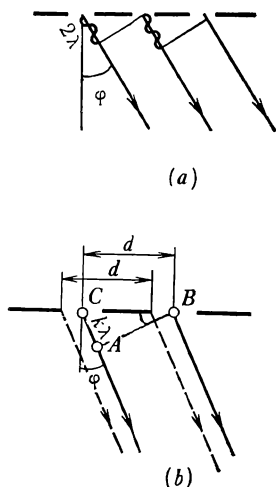


Fig. 35.16 Bright fringes on screen are produced by rays whose path difference are equal to integral number of wavelengths.



fication of waves propagating only in certain directions and several narrow bright fringes are formed on the screen.

How can we find the directions of maximum intensification of light of wavelength  $\lambda$ ? Figure 35.15b shows rays making an angle  $\phi$  with the normal to the grating. This is the direction for which the path difference for rays passing through adjacent slits is equal to one full wavelength. All the rays going in this direction are collected by a lens at point  $M$ , at which point a bright spot is visible on the screen.

It is obvious that the angle  $\phi$  can be increased until another direction is found for which the path difference for rays passing through adjacent slits is two wavelengths (Fig. 35.16a), three wavelengths, etc. Thus, the directions in which rays forming bright fringes on the screen propagate are distinguished by the property that the path differences for rays passing through adjacent slits are always equal to  $k\lambda$ , where  $k$  is an integer. In Fig. 35.16b,  $AB$  is seen to be the wavefront ( $AB \perp AC$ ), the path difference for the rays  $AC$  being  $k\lambda$ . Since the angle  $B$  in  $\triangle ABC$  is  $\phi$  and the hypotenuse  $BC = d$ , we have the formula for a diffraction grating

$$k\lambda = d \sin \phi \quad (35.5)$$

This formula holds also for  $k = 0$ , since there is a bright line on the screen opposite to the centre of the grating. The term for each bright line on the screen is *maximum* and for the number  $k$  corresponding to each maximum its *order*. Thus, the brightest maximum for the zeroth order is visible on the screen opposite the centre of the grating; at equal distances on both sides of it the less bright first maxima are visible followed by the maxima of the second order, still less bright, etc. Experiment shows all these maxima to be equidistant from one another (Fig. 35.17a).

Let monochromatic light of a greater wavelength fall on the same grating. In this case the maxima will be more sparse (Fig. 35.17b), but the zeroth maximum will be in the same part of the screen for both  $\lambda$ 's. Hence, the position of the zeroth maximum is independent of  $\lambda$ , that is, it remains the same for all wavelengths.

It follows from formula (35.5) that to measure the length of a light wave the only thing one has to do is to measure the angle  $\phi$ , since  $d$  and  $k$  are always known. Since angles can be measured with a great degree of accuracy, the wavelength  $\lambda$  can also be measured with a great degree of accuracy. Note that the accuracy of measurement of  $\lambda$  increases with the decrease in the spacing  $d$  of the grating.

The diffraction grating can be used to determine the composition of luminous radiation, since light of different wave-

lengths produces maxima in different parts of the screen. An increase in the total number of rulings in the grating makes the fringes on the screen thinner, thus making it possible to see separate maxima of rays of smaller differences in wavelength. This means that the increase in the total number of rulings in the grating increases its *resolving power*.

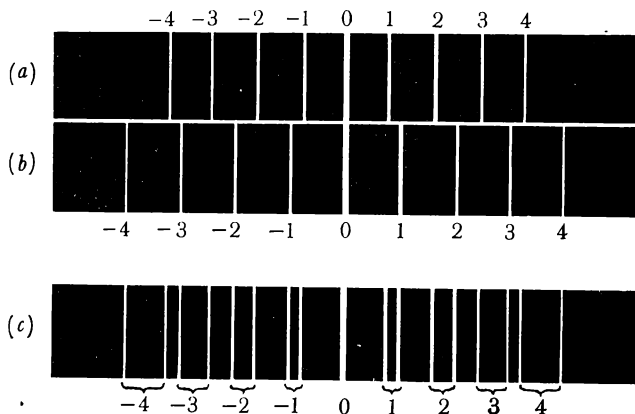


Fig. 35.17 Pattern seen on screen when diffraction grating is illuminated with parallel (a) violet rays, (b) blue rays, (c) blue and violet rays.

For a constant difference in wavelength the spacing between the maxima in Fig. 35.17a is seen to increase with their order,  $k$ . Therefore the increase in  $k$  results in a higher resolving power, but also poorer visibility of the pattern on the screen.

The distribution of radiation over frequencies (over wavelengths) is termed *spectrum of radiation* (from the Latin word for visible). By letting white light fall on a grating one can obtain its spectrum. As has been already stated, it consists of all the colours present in a rainbow (see Plate (c) in the colour insert). The maxima on both sides of the white line, 0, are arranged in order of increasing wavelength, the distance between the coloured lines in the same spectrum being proportional to the difference in their wavelength, that is, the diffraction spectrum is uniformly extended in all its parts (this is said to be the *normal spectrum*). The insert shows that among the visible rays violet rays have the shortest wavelength and red ones the longest. Note that spectra of great values of  $k$  may overlap, and this complicates their observation.

An interesting picture can be observed on the screen if two identical crossed gratings (i.e. with their slits at right angles) are placed in the path of the light rays. This picture is presented in Fig. 35.18. It consists of individual bright

Fig. 35.18 Bright spots on the screen obtained from two crossed gratings.

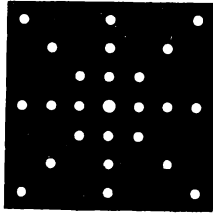
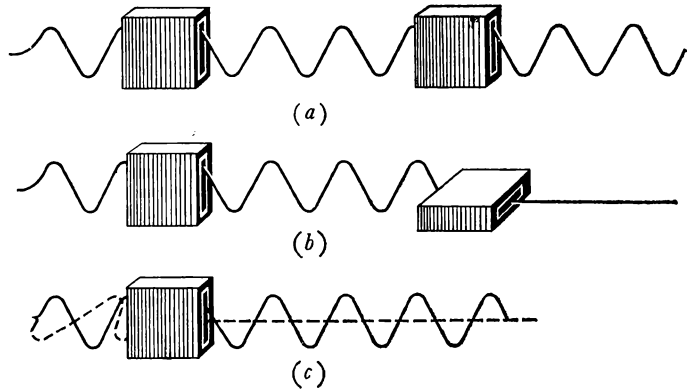


Fig. 35.19 (a) Transverse wave travelling along cord freely passes through boxes when their slits are parallel; (b) second box stops wave when its slit is at right angles to oscillations of cord; (c) box passes only such waves whose oscillations are parallel to its slit.



the wave will travel through the box only if the oscillations in it coincide in direction with the slits.

Place two such boxes one behind the other in the direction of propagation of transverse waves. If the boxes are arranged with their slits parallel, the transverse wave will pass through both boxes and propagate further only if its oscillations are in the direction of the slits in the boxes (Fig. 35.19a). If the boxes are crossed (i.e. their slits are arranged at right angles (Fig. 35.19b)), there is no possibility of the transverse wave passing through them. Obviously, all this refers only to transverse waves, since a longitudinal wave will pass through the slits in the boxes no matter what their mutual arrangement.

Hence, in the arrangement of two boxes of the type described above, one placed behind the other along the direction

spots. When the diffraction gratings are of different spacings and are not in contact, the pattern is a more intricate one. Actually, an analysis of the disposition of the spots on the screen makes it possible to find the distance between the gratings and their spacings. This opened the way for the determination of the arrangement of atoms in the lattices of solids (see Section 37-15).

### 35-7 Polarization of Waves

The third proof of the wave nature of light is the phenomenon of *polarization*. This is possible only for transverse waves. The easiest way to uncover the essence of this phenomenon is to use the example of mechanical waves.

Let a transverse wave run along a cord through a narrow box with slits in the opposite faces (Fig. 35.19a). Obviously,

of wave propagation, the rotation of one box about this direction can always stop a transverse wave although a longitudinal wave cannot be stopped. At this point one is entitled to ask the question: Why are there two boxes? In fact, if there is only one box and it is rotated about the direction of wave propagation, a position can be found (when the slit is at right angles to the direction of oscillations in the wave) when it will stop the transverse wave.

To answer the question draw two cords through the box and excite oscillations in them in perpendicular directions (Fig. 35.19c). The box will stop the wave from propagating only along one cord, the wave along the other being able to pass through it. One more box is needed to stop the other wave.

Hence, if there are transverse waves with oscillations simultaneously taking place along mutually perpendicular directions lying in the plane normal to the wave propagation direction, they can be stopped completely only with the aid of two boxes. The first box will transmit only the wave whose oscillations coincide in direction with its slit, and this wave can be stopped by the second box appropriately rotated about the direction of its propagation.

If the directions of the oscillations of all the points of a ray of a transverse wave lie in the same plane, the wave is said to be *plane polarized*, the term for the plane normal to the direction of the oscillations being *polarization plane*. Figure 35.19a and b depicts plane polarized waves for which the polarization plane is normal to the plane of the figure. Note that a plane polarized wave can be stopped by one box. The term for the device used to find out whether the wave passing through it is polarized is *analyzer* (the second box). The term for the device which makes an unpolarized wave passing through it polarized is *polarizer* (the first box).

It follows from the above description that the polarizer and the analyzer are identical in design. The same device may be used either as a polarizer or as an analyzer, as the circumstances demand. We would like to remind the reader once again that the concept of polarization is meaningless for longitudinal waves. Only transverse waves can be polarized.

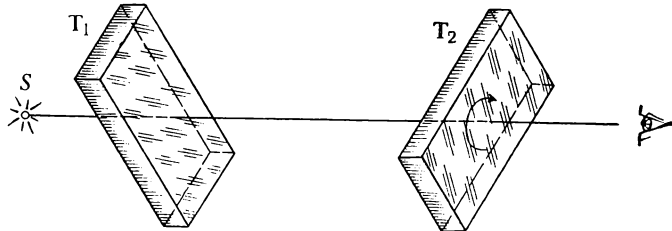
### 35-8 Polarization of Light

It has been established by experiment that light rays can be polarized. The first polarizers of luminous radiation were made of single crystals of tourmaline (a natural mineral).

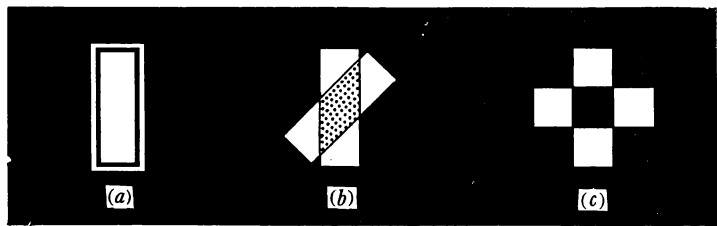
Plates identically oriented with respect to the crystal axes (see Sections 13-1 and 13-2) are cut out of such crystals and placed one behind the other in the path of a light ray (Fig. 35.20).

The plate  $T_1$  acts as a polarizer and the plate  $T_2$  as an analyzer. When one rotates the analyzer one first observes that the intensity of light diminishes to a minimum (with

**Fig. 35.20** Tourmaline plates for observation of polarization of light rays.

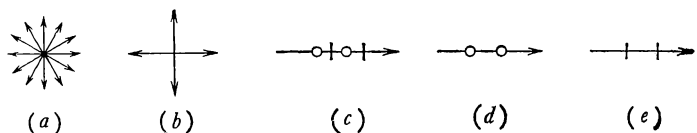


**Fig. 35.21** Patterns seen when one tourmaline plate is rotated; the plates are (a) parallel, (b) at an oblique angle and (c) crossed.



the plates crossed) and then increases again (Fig. 35.21). If the polarizer is withdrawn, the analyzer's rotation fails to change the intensity of the light reaching the observer's eye.

**Fig. 35.22** Schematic representation of cross section of (a) natural ray and (b) model ray; (c) schematic representation of the model ray: circles correspond to oscillations normal to plane of figure and dashes to those lying in its plane.



To sum up, experiments carried out with luminous radiation have demonstrated that

(1) luminous radiation propagates in the form of transverse waves;

(2) the oscillations in a ray of natural light occur in all directions lying in the plane normal to it, with no one direction being dominant.

To describe the polarization phenomenon in practice a model ray is used, the oscillations in which occur along two mutually perpendicular directions of arbitrary choice (Fig. 35.22).

Let us now try to understand why tourmaline acts as a polarizer of light. Recall that crystals are anisotropic (Section 13-2). Tourmaline's anisotropy manifests itself in its ability to intensely absorb radiation with oscillations taking place along one definite direction and to transmit radiation with oscillations in the direction perpendicular to the former practically without absorption. The term for this property of crystals is *dichroism*.

Using a tourmaline plate thick enough for one polarized ray to be completely absorbed, one obtains a *completely polarized ray*. If the plate is thinner, the transmitted ray retains oscillations in mutually perpendicular directions, but the amplitude in one direction is greater than in the other. Such a ray is said to be *partially polarized*.

A very pronounced dichroism was discovered in tiny quinine iodide sulphate crystals. In the process of fabrication of the polarizer a celluloid film is coated with a thin layer of such crystals oriented in a special way. This film is covered with a glass plate and in this way a large-scale polarizer is obtained. The term for such a polarizer is *polaroid*.

Some materials have been found to possess very interesting properties. When a polarized ray enters them, its polarization plane rotates in proportion to the distance travelled by the ray in the material. The term for materials capable of rotating the polarization plane of a ray is *optically active*. They include quartz, a solution of sugar in water, etc.

A good proof of the electromagnetic nature of luminous radiation is the magnetic rotation of the polarization plane of a ray in a magnetic field discovered by Faraday: when a polarized ray propagates in the direction of magnetic induction lines, its polarization plane rotates.

### 35-9 Polarization of Light by Reflection and Refraction

We recall that in a natural ray the vector  $\mathbf{E}$  oscillates in all directions in a plane normal to the ray. It was established that as the result of reflection and refraction of rays at a boundary separating two transparent media the reflected and refracted rays experience partial polarization.

Figure 35.23a is a schematic diagram of a model ray falling at an angle of incidence  $i$  on a boundary separating air and a liquid. The reflected ray abounds in oscillations parallel to the boundary surface (denoted by circles); the refracted ray in oscillations normal to them (denoted by

Fig. 35.23 (a) Polarization by reflection and refraction: reflected ray is partly polarized in plane of figure (more circles than dashes) and refracted ray in plane normal to figure (more dashes than circles); (b) reflected ray is completely polarized when it makes  $90^\circ$  with refracted ray.

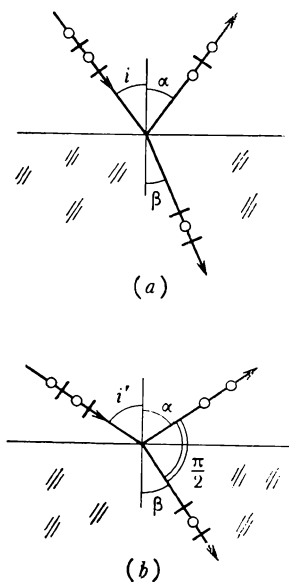
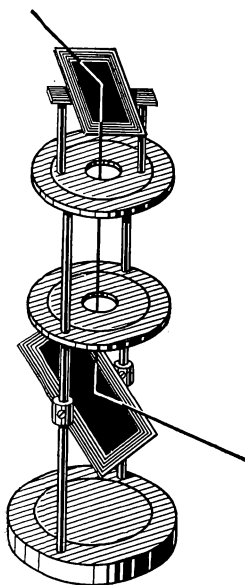




Fig. 35.24 Observation of polarization by reflection.



dashes). The degree of polarization of these rays depends on the angle of incidence  $i$  and on the refractive index  $n$ .

It was established as a result of studies of this phenomenon that for transparent materials the polarization of the refracted ray is always partial and that for the reflected ray there is one direction in which it is completely polarized (Fig. 35.23*b*). Actually, complete polarization of the reflected ray takes place when the angle between the reflected and the refracted ray is  $\pi/2$ . Hence

$$\alpha + \beta = \pi/2 \quad \text{and} \quad \beta = \pi/2 - \alpha$$

We denote the incidence angle for this case by  $i'$ . Taking into account that  $\angle \alpha = \angle i'$ , we obtain from the second law of refraction

$$\frac{\sin i'}{\sin(\pi/2 - i')} = \frac{\sin i'}{\cos i'} = \tan i' = n$$

This gives us *Brewster's law*: the tangent of the angle of incidence corresponding to complete polarization ( $i'$ ) is equal to the relative refractive index. Note that the degree of polarization of the refracted ray is maximum at the angle of incidence  $i'$  as compared with other angles of incidence.

The polarization of rays as the result of reflection is sometimes used to devise polarizers and analyzers. Such a device consists of two glass plates (Fig. 35.24) with blackened lower surfaces to absorb refracted rays. One of the plates, for instance, the lower, acts as a polarizer. It is arranged so as to make the light ray falling on it completely polarized and to direct it onto the second mirror.

Rotating the second plate about both the vertical and horizontal axes, one can find a position of the plate at which the reflected ray vanishes completely. Thus the second plate acts as an analyzer.

## 36

## Photometry

### 36-1 Energy Flux of Radiation. Solid Angle

Electromagnetic radiation travelling in a medium, like all waves, transports energy from point to point. Imagine a surface through which waves pass at some distance from the source of the electromagnetic radiation. The energy transported through this surface per unit time is termed the *energy flux of radiation* through this surface; it has the dimension of power and is measured in watts.

When the distance from the source of the electromagnetic radiation to the chosen surface is large compared to the dimensions of the source itself, the source can be said to be a *point source*. It is often assumed that the radiation of a point source is independent of direction, that is, the point radiates uniformly in all directions.

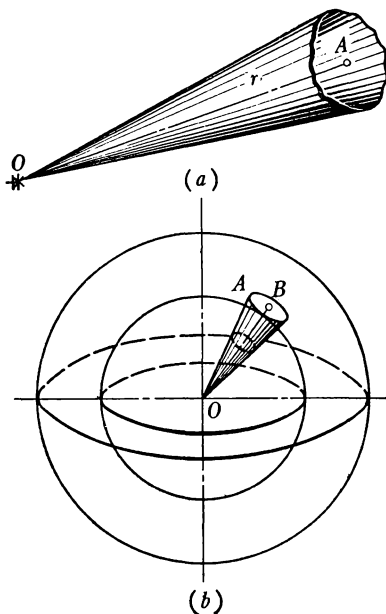


Fig. 36.1 (a) Solid angle; (b) central solid angle.

The flux of radiation falling on a surface depends on the area of this surface  $A$ , on its orientation in space and on its distance from the source of the radiation.

In most cases one comes up against fluxes propagating in a limited region of space. For instance, if a source of radiation  $O$ , small in size compared to  $r$  (Fig. 36.1a), sends its radiation to an area  $A$  normal to the direction of radiation, the latter receives only the radiation bounded by the shaded conical surface with its apex at point  $O$ .

The part of space bounded by the conical surface is termed *solid angle*,  $\Omega$ . The term for point  $O$  in Fig. 36.1 is *apex* of solid angle. When the apex is in the centre of a sphere, the angle is termed *central*. If spherical surfaces of different radii  $r$  with their centres at point  $O$  (Fig. 36.1b) are drawn, then for a specified solid angle the ratio of  $A$  to  $r^2$ , as we know from geometry, will be the same for all surfaces and can serve as a measure of the solid angle  $\Omega$  that cuts an area  $A$  out of

a spherical surface:

$$\Omega = \frac{A}{r^2} \quad (36.1)$$

We find the unit for measuring  $\Omega$ :

$$\Omega = \frac{1 \text{ m}^2}{1 \text{ m}^2} = 1 \text{ sr (steradian)}$$

*Steradian* is the term for a central solid angle that cuts out of a spherical surface an area equal to the square of the sphere's radius. Since the formula for the area of a sphere is  $A_{\text{sp}} = 4\pi r^2$ , the spherical surface contains  $4\pi$  areas equal to the square of its radius. This means that the total solid angle covering the entire space has  $4\pi$  sr:

$$\Omega_{\text{total}} = 4\pi \text{ sr} \quad (36.2)$$

When one wants to measure the radiation propagating from the source  $O$  in a specified direction  $OB$  (see Fig. 36.1*b*), one considers the flux of radiation inside a very small angle  $\Omega$  which cuts out of the spherical surface an area  $A$  with its centre at  $B$ . Dividing the whole surface area of the sphere into equal areas  $A$  and measuring the fluxes passing through each of them one can find in which direction the flux is at its maximum and in which it is at its minimum.

### 36-2 Luminous Flux

The radiation responsible for a person's sense of vision is electromagnetic radiation in the range of wavelengths from 400 to 760 nm (in a vacuum), each frequency band inside this range corresponding to a different sense of colour (see Section 31-2).

It has been demonstrated in experiments that light fluxes of equal intensity but different wavelengths cause different levels of stimulation of the terminations of the optic nerve in the eye's retina and, accordingly, create visual sensations not only of different colour but also of different intensity. Our eyes are most sensitive to radiation with a wavelength of 555 nm (green light). Identical fluxes of shorter or longer wavelengths excite visual sensations of lesser intensity. For the purpose of a qualitative assessment of this difference we adopt the following procedure.

Take sources of monochromatic radiation of different colours but equal power (say, one watt) and compare them in similar conditions in turn with a variable power source of 555-nm radiation. We can then adjust the power of the standard source with the wavelength  $\lambda = 555 \text{ nm}$  so that it

creates visual sensations equal in intensity to those created by each individual source. One possible way of comparing the sources is, for instance, to use them to illuminate adjacent areas of the same text to make them equally clear (readable).

Let us introduce the term *relative visibility factor* for the ratio of the power of the standard source with a wavelength of 555 nm to the power of the monochromatic source being compared with it. We find that, for instance, a flux of orange light ( $\lambda = 610$  nm) with a power of 1 W creates a visual sensation equal in intensity to that created by a flux of green light ( $\lambda = 555$  nm) of 0.5 W. This means that for  $\lambda = 610$  nm the relative visibility factor  $K = 0.5$ .

The values of the relative visibility factor for various colours are presented in Table 36.1. (Obviously, for  $\lambda = 555$  nm  $K = 1$ .) Figure 36.2 depicts the plot of the relative visibility factor against wavelength (in a vacuum). The term for this plot is *relative visibility curve*, or the *graph of spectral sensitivity* of the eye. Note that at nighttime this curve shifts a little in the direction of shorter wavelengths, that is, to the left.

Fig. 36.2. Relative visibility curve.

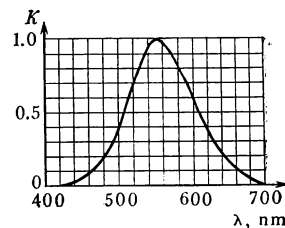


Table 36.1 Relative visibility coefficients

Spectral band	Wavelength, nm	Coefficient of relative visibility
Violet	{ 400	0.0004
	{ 440	0.023
Blue	{ 470	0.091
	{ 495	0.265
Green	{ 555	1
	{ 580	0.837
Yellow and orange	{ 610	0.503
	{ 640	0.175
Red	{ 700	0.0041
	{ 750	0.0001

The conclusions to be drawn from the above is that it is very inconvenient to express the light flux which creates visual sensation in the human eye in watts. For this reason *luminous flux*  $\Phi$  is used as a measure of the effect of radiation on the human eye. The term *luminous flux* applies to that part of radiation flux which creates visual sensation in the eye and which is measured on the basis of this sensation.

The area of optics which deals with the measurement of luminous flux and with the study of the characteristics of light sources and the illuminance of objects is termed *photometry* (from the Greek *photos* for light). Note in addition that the part of electromagnetic radiation responsible for visual sensations is often termed *luminous radiation*, the luminous flux  $\Phi$  being the measure of it.

### 36-3 Luminous Intensity

A luminous flux  $\Phi$  is always the product of some light source. The radiation of real light sources is never uniform in all directions.

The term for the quantity characterizing the dependence of the luminous flux on the direction of radiation is *luminous intensity*,  $J$ . The measure of the luminous intensity of a light source of small dimensions is the luminous flux radiated by this source in a unit solid angle pointing in the specified direction:

$$J = \frac{\Phi}{\Omega} \quad (36.3)$$

To measure the luminous intensity of a real light source in a given direction one measures the luminous flux  $\Phi$  inside a small angle  $\Omega$  and then uses formula (36.3) to find  $J$ . When the dependence of the luminous intensity of a light source on direction is not strong, formula (36.3) is valid for large  $\Omega$ 's. In the following we assume the luminous intensity of a point source to be independent of direction.

The unit of luminous intensity in the SI system, the *candela* (from the Latin word for candle), is the sixth base unit. The term candela (cd) applies to 1/60 of the luminous intensity radiated by 1 cm<sup>2</sup> of a plane platinum surface\* at the temperature of its solidification (2046 K), in a direction normal to the surface and at a pressure of 1 atm.

A parameter sometimes used to describe light sources with a marked dependence of luminous intensity on direction is *average spherical luminous intensity*  $J_{av}$ . It is obtained from the relation

$$J_{av} = \frac{\Phi_{total}}{4\pi} \quad (36.4)$$

where  $\Phi_{total}$  is the total luminous flux of the source.

\* To be more exact, by 1 cm<sup>2</sup> of the surface of a black body (see Section 37-10) at the temperature of the solidification of platinum.

We deduce a unit for measuring luminous flux:

$$\Phi = J\Omega, \quad \Phi = 1 \text{ cd} \times 1 \text{ sr} = 1 \text{ lm (lumen)}$$

The unit for measuring luminous flux in the SI system is the *lumen*. The term lumen applies to the luminous flux radiated by a point light source of 1 cd into a solid angle of 1 sr.

Since the total solid angle contains  $4\pi$  steradians, the total flux radiated by a point source will be expressed by the formula

$$\Phi_{\text{total}} = 4\pi J \quad (36.5)$$

It was established as the result of measurements that 1 lm of a monochromatic flux of 555-nm wavelength corresponds to a flux of radiation of 0.001 61 W, that is, 1 W of such radiation is equal to 621 lm.

The term for the ratio of the number of lumens in the luminous flux  $\Phi$  of an electric light source to the power consumed by the source in watts  $P$  is *luminous efficiency*  $k$  of the source:

$$k = \frac{\Phi}{P} \quad (36.6)$$

Table 36.2 gives the values of the luminous efficiencies of electric bulbs of different power.

**Table 36.2** Luminous characteristics of incandescent lamps

Power, W	Total luminous flux, lm	Luminous efficiency, lm/W	Average spherical luminous intensity, cd
15	124	8.25	10
25	225	9.00	18
40	380	9.50	30
60	645	10.75	51
100	1275	12.75	103
150	2175	14.50	173
200	3050	15.25	243
300	4875	16.25	388
500	8725	17.45	695
1000	19000	19.00	1530

### 36-4 Illuminance

During a dark night or in a cave the objects around us are invisible. But a burning match, and objects close to it, are clearly visible in such circumstances. This is because a lumi-

nous flux propagates from a light source, in this case from a burning match. Part of the luminous flux, falling on objects, is scattered by them and reaches a person's eyes, enabling him to see them. The greater the luminous flux falling on such objects, the greater the scattered flux and the more clearly visible the objects become.

The term for the quantity  $E$  characterizing the visibility of various bodies due to the luminous fluxes falling on them is *illuminance*. For a uniform distribution of the luminous flux over the surface of the body the measure for its illuminance is the luminous flux falling on a unit area of this surface:

$$E = \frac{\Phi}{A} \quad (36.7)$$

When the illuminance of different areas of the surface is not the same, one should choose an area  $A$  small enough, so that the distribution of the luminous flux over it can be assumed uniform. When formula (36.7) is used to estimate illuminance in the case of a nonuniform distribution of the flux  $\Phi$  over the area  $A$ , the result will be the average illuminance of this surface.

Deduce a unit for measuring the illuminance  $E$ :

$$E = 1 \text{ lm}/1 \text{ m}^2 = 1 \text{ lm}/\text{m}^2 = 1 \text{ lx (lux)}$$

The unit used for measuring illuminance in the SI system is the *lux* (from the Latin word for light). The term lux applies to the illuminance of a surface each square metre of which receives a luminous flux of one lumen. Some typical illuminance values and artificial lighting standards are presented in Table 36.3.

**Table 36.3** Characteristic values of illuminance

Illuminance	$E$ , lux
Sun rays at midday (at medium latitudes)	100000
In a film studio	10000
On an open spot on a cloudy day	1000
At a working place for fine work (drawing)	100-200
In a well-lighted room (close to the window)	100
For reading in classes and laboratories	up to 75
On a cinema screen	20-80
In corridors and on stair-cases	up to 15
On artificially lighted streets	up to 4
From a full Moon	0.2
From stars in the sky on a moonless night	0.0003

### 36-5 Luminance

Reading a book or a note written on paper we clearly see letters on the white background of the page despite the fact that the page is uniformly illuminated. This is because the white page and the letters scatter the luminous flux falling on them in a different way.

Since the paper page radiates luminous flux, it can be regarded as a light source. Note that the page radiates not its own light but scattered light; therefore it is conveniently termed a *secondary light source*. The magnitude of luminous flux propagating from both a primary and a secondary light source generally depends on direction. This means that, like primary light sources, secondary light sources can be characterized by luminous intensity. The white surface of the page appears to us much brighter than the letters written on it. Therefore the luminous intensity in the former case should be greater than in the latter.

Thus, different areas of surfaces of real light sources (both primary and secondary) viewed from a specified direction may have greatly differing luminances, for instance some turns of an electric radiator connected to the main may appear much brighter than the others.

The term for the quantity  $B$  characterizing the visibility of different parts of a surface viewed from a specified direction due to the luminous flux radiated by this surface is *luminance*. For a uniform distribution of the flux radiated in a specified direction over a surface the measure of luminance is luminous intensity radiated by a unit area of the surface. If the luminous intensity is measured along the normal to the surface, its luminance is determined from the formula

$$B = J/A \quad (36.8)$$

Deduce a unit measuring luminance:

$$B = 1 \text{ cd}/1 \text{ m}^2 = 1 \text{ cd}/\text{m}^2 = 1 \text{ nt (nit)}$$

The unit of luminance in the SI system is  $\text{cd}/\text{m}^2$  (or *nit*)—the luminance of a plane surface each square metre of which radiates a flux of luminous intensity of 1 cd in the direction normal to it.

Note that the minimum luminance felt by the human eye is about  $10^{-6}$  nt, while a luminance in excess of  $10^5$  nt causes pain and possible harm to the eye. Table 36.4 shows typical values of luminance for various surfaces.



Table 36.4 Luminance of some light sources and surfaces

Source, surface	$B$ , nit
Surface of the Sun	$1.5 \times 10^9$
Electrode crater in electric arc	$1.5 \times 10^8$
Metal filament of incandescent lamp	$(1.5-2) \times 10^6$
Snow illuminated by direct sunlight	$3 \times 10^4$
Flame of a kerosene lamp	$1.5 \times 10^4$
Flame of a candle	$5 \times 10^3$
Clear blue sky	$4 \times 10^3$
Surface of the Moon	$2.5 \times 10^3$
Surface of a cinema screen	5-20
Sheet of white paper (illuminance 30-50 lux)	10-15
Clear moonless night sky	$10^{-4}$

### 36-6 The Laws of Lumination

The illuminance on a surface produced by a point light source depends on the luminous intensity  $J$  and on the distance to the surface  $r$ .

Draw a sphere of radius  $r$  around a point source with a luminous intensity  $J$ . In such a case the illuminance will be identical everywhere on the inside surface of the sphere, the direction of the rays coinciding with the radii (i.e. being normal to the sphere). Hence, the angle of incidence of the rays on the sphere is zero. Introducing the notations  $E_0$  for the illuminance of the inside surface of the sphere in these conditions,  $A_{sp}$  for the total area of the inside surface, and  $\Phi_{total}$  for the total luminous flux from the source, we obtain from (36.7)

$$E_0 = \frac{\Phi_{total}}{A_{sp}}$$

Since  $\Phi_{total} = 4\pi J$  and  $A_{sp} = 4\pi r^2$ , it follows that  $E_0 = \frac{4\pi J}{4\pi r^2}$ , or

$$E_0 = J/r^2 \quad (36.9)$$

This relation is a mathematical expression for the *first law of illuminance*: in the case of the normal incidence of rays, the illuminance produced by a point light source is directly proportional to its luminous intensity and inversely proportional to the square of the distance from the source to the surface illuminated by it.

Consider now the dependence of illuminance on the angle of incidence of the rays. Let parallel rays of light strike the

plane surface  $FBCD$  at an angle of incidence  $i$  (Fig. 36.3). The formula for the illuminance  $E$  of this surface is

$$E = \frac{\Phi}{A} = \frac{\Phi}{FB \times FD}$$

where  $\Phi$  is the luminous flux falling on the surface  $FBCD$ .

If the surface  $FBCD$  is withdrawn, the luminous flux will fall on the surface  $MNCD$ . Let this surface be such that the angle of incidence of rays falling on it is zero. In this case the angle between the surfaces  $FBCD$  and  $MNCD$  is  $i$ . Denote the illuminance of the surface  $MNCD$  by  $E_0$ ; then

$$E_0 = \frac{\Phi}{MN \times MD}$$

Find the ratio of illuminance  $E$  and  $E_0$ :

$$\frac{E}{E_0} = \frac{\Phi \times MN \times MD}{\Phi \times FB \times FD}$$

Since  $FB = MN$ , it follows that

$$E/E_0 = MD/FD = \cos i$$

Hence

$$E = E_0 \cos i \quad (36.10)$$

This relation is a mathematical expression for the *second law of illuminance*: the illuminance of a surface produced by parallel rays is directly proportional to the cosine of the angle of incidence of the rays on this surface.

It follows from the second law of illuminance that an increase in the angle of incidence results in a decrease in illuminance. The explanation for the change of seasons on the Earth is just such change in the angle of incidence of the solar rays striking the ground. In the Northern Hemisphere the minimum angle of incidence of solar rays on the ground is at the end of June and the maximum at the end of December. (Explain why in the Southern Hemisphere December is a summer month and June a winter month.)

For a point light source for  $E_0$  in formula (36.10) we can substitute the value obtained from (36.9). We can then obtain a generalized formula for calculating illuminance:

$$E = J/r^2 \cos i \quad (36.11)$$

This formula can be used to compute the illuminance of various parts of a surface (for instance, different points on a table) produced by an electric bulb (Fig. 36.4). In the course of the calculations one should keep in mind that the illuminance produced by various light sources at some point of a surface is the sum of the illuminances produced by each individual source.

Fig. 36.3 Dependence of illuminance on angle of incidence of parallel rays.

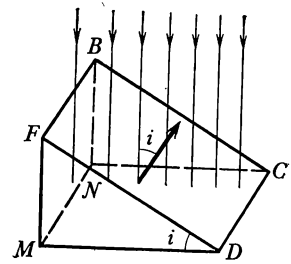
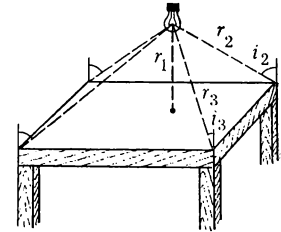


Fig. 36.4 Illuminance depends on distance to light source and on angle of incidence.



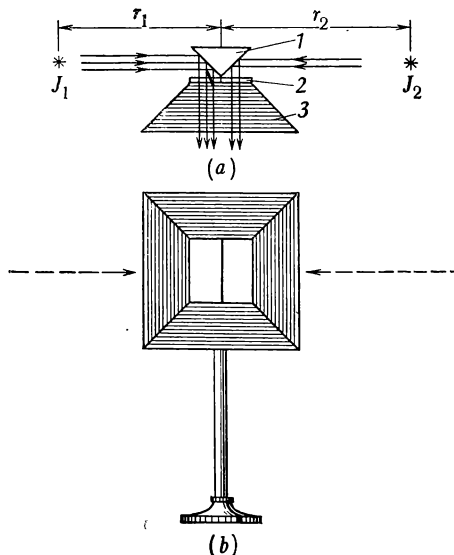
### 36-7 Light Measurements

If one knows the luminous intensity of a light source he can easily find the other photometric quantities. Luminous intensity is measured with an instrument called the *photometer*. In the photometer of the simplest type the method used to measure an unknown luminous intensity is to compare it with a known luminous intensity of a standard source by obtaining equal illuminances from both sources.

Note that a person is able to establish the equality of the illuminances of adjacent surfaces very precisely. However, in the case of different illuminances he is unable to estimate their ratio.

One of the photometers is shown schematically in Fig. 36.5. It consists of a triangular prism *1* coated with white paint

Fig. 36.5 Photometer:  
(a) view from above; (b)  
front view.



capable of effectively scattering light. The light sources are placed to the left and to the right of the prism. The rays scattered by the prism fall on mat glass 2 and pass inside a shade 3 to reach the eye. One of the light sources, for instance the standard one, is placed at a definite distance  $r_1$  from the photometer, and the second source is shifted to the left or to the right until the illuminances of both halves of glass 2 are equal. After that the distance  $r_2$  from the second source to the photometer is measured. In compliance with the first law of illuminance

$$E_1 = J_1/r_1^2 \quad \text{and} \quad E_2 = J_2/r_2^2$$

Since the illuminances  $E_1$  and  $E_2$  are equal, we have

$$J_1/r_1^2 = J_2/r_2^2$$

whence

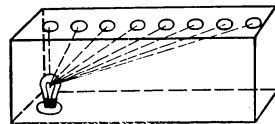
$$J_1/J_2 = r_1^2/r_2^2 \quad (36.12)$$

From this formula we can find the unknown luminous intensity  $J_2$ .

The instrument used for measuring the magnitude of illuminance is the *luxmeter*. A luxmeter of the simplest type is a box with an electric bulb placed at one of its sides (Fig. 36.6). The lid of the box has holes covered with white paper. The farther the hole from the bulb the less its illuminance. Opposite each hole there is a number indicating its illuminance in luxes. To find the illuminance of the table the luxmeter (with the bulb burning) is placed on it and the hole is found whose pattern on the white paper covering the box is undistinguishable from the background. The illuminance on the table is equal to the one indicated opposite the hole. (Why?) All other hole patterns will appear as bright circles on a dark background or as dark circles on a bright background.

The photometers and luxmeters of the simplest types described above are not very accurate. Photometers and luxmeters of advanced types employ photocells (see Section 38-13). An example of an instrument of this sort is the exposure meter used to measure illuminance and to determine exposure time in photography.

Fig. 36.6 Luxmeter.



## Radiation and Spectra. X Rays

## 37

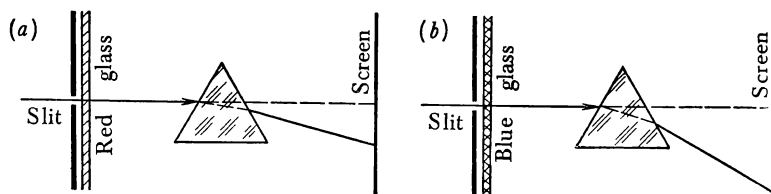
### 37-1 Dispersion of Light

Take a triangular glass prism and place it between a light source in the shape of a slit and a screen. Directing first a red beam on the prism (Fig. 37.1a) and then a blue one (Fig. 37.1b), we notice that the blue beam passing through the prism is deflected from its original direction more than the red one. This means that the absolute refractive index of glass is less for red rays ( $n_r$ ) than for blue rays ( $n_b$ ). Since  $n_b = c/v_b$  and  $n_r = c/v_r$ , we have

$$c/v_b > c/v_r, \text{ or } v_r > v_b$$

Hence, the velocity of propagation of red rays in glass is greater than of blue rays. The propagation velocity of light rays in glass is the less the greater their frequency, or the less their wavelength. A similar dependence is observed in other transparent materials as well.

Fig. 37.1 Refraction of monochromatic rays in prism: blue ray is refracted more than red.

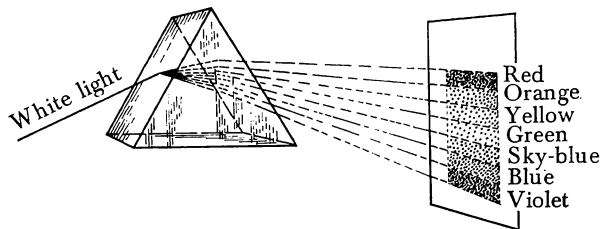


The term for the dependence of the velocity of propagation of waves in a medium on their wavelength (frequency) is *dispersion*. In practice dispersion is expressed in terms of the dependence of the refractive index on the frequency or wavelength in the material. It is an experimental fact that in the great majority of cases the refractive index decreases with the wavelength. This kind of dispersion is termed *normal*.

### 37-2 Dispersion by a Prism

Using a triangular glass prism, Sir Isaac Newton first discovered in 1666 that white light has a continuous spectrum (see Section 35-6). A remarkable feature of white light is

Fig. 37.2 Decomposition of white light by prism (spectrum of white light is seen on screen).



that the spectrum is made up of a continuous succession of monochromatic rays. For this reason such a spectrum is termed *continuous*.

Newton divided the continuous spectrum of white light into seven bands, each of a separate colour: red, orange, yellow, green, sky-blue, blue and violet. As they emerge from the prism, the colours arrange themselves in order of decreasing wavelength (Fig. 37.2). We recall that the spect-

rum of white light can also be obtained with the aid of a diffraction grating. The latter spectrum is termed a *diffraction*, or *normal*, *spectrum*. The dispersion curve for glass (Fig. 37.3) shows that in the short-wave range the refractive index of glass changes rapidly with the wavelength, its

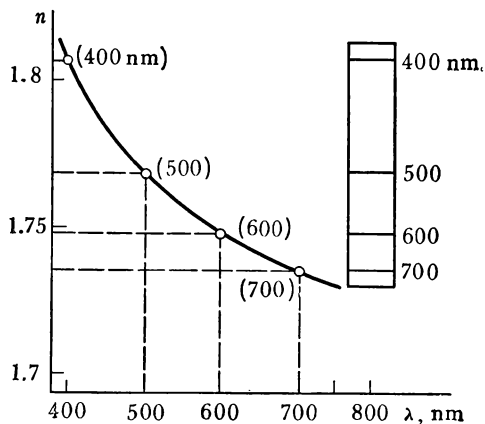


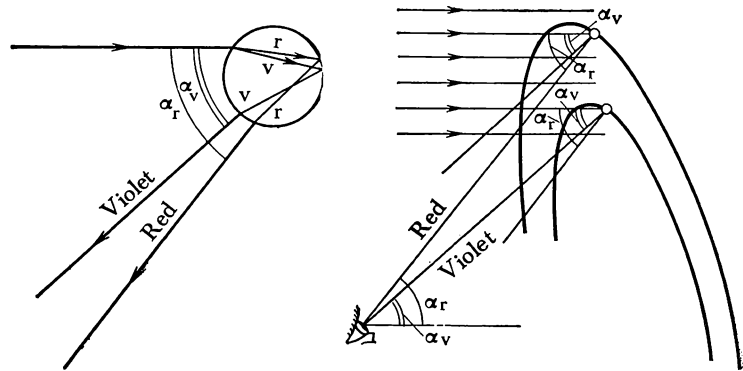
Fig. 37.3 Dispersion curve for glass.

change in the long-wave range being much less pronounced. Because of this the dispersion (prismatic) spectrum of white light is compressed in the red part and extended in the violet. The difference between the prismatic (dispersion) spectrum and the normal spectrum of white light (see the colour insert) is that the latter, firstly, consists of colour bands arranged in order of increasing wavelengths and, secondly, that the bands in it are uniformly distributed in all its parts (see Section 35-6).

Dispersion of light is what causes the *rainbow*. A rainbow is visible when the observer has the Sun behind him and when there are water droplets suspended in the air. At a definite angle of incidence there is total reflection of light inside a droplet (Fig. 37.4). The rays are refracted at the boundary separating the air from the water and, since violet rays are refracted more than red ones, both rays diverge as they emerge from the droplet: the red rays make an angle of about  $43^\circ$  with the angle of incidence and the violet about  $41^\circ$  (see Fig. 37.4).

Solar rays can be assumed to be parallel. Therefore the observer's eye receives red rays emerging from numerous droplets arranged on the surface of a cone with an apex angle of  $\alpha_r = 43^\circ$  and violet rays emerging from droplets

Fig. 37.4 Formation of rainbow.

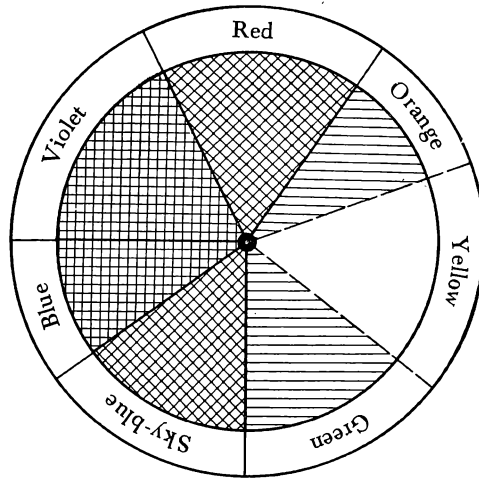


arranged on the surface of another cone with an apex angle of  $\alpha_v = 41^\circ$ . The other colours of the rainbow are in between.

### 37-3 Combining Colours. Complementary Colours

If the colour rays of which the white light is composed are combined to form one ray, the result will again be white light.

Fig. 37.5 Newton's disk; rotating disk appears to be coloured gray.



Such a mixture of colours can be made with the aid of Newton's circle (Fig. 37.5). Its sections are painted in the seven colours of the rainbow, and when man looks at the rotating disk, his eyes alternately receive the rays of

all the seven colours, scattered by the coloured sections of the disk. Since the persistence of vision of the human eye is about 0.1 s, a person sees the quickly spinning disk as grey. The fact that the colour is grey instead of white is due to the absence of intermediate colours and the low quality of the paint.

The colours corresponding to monochromatic radiation are sometimes termed *spectral*. A mixture of two monochromatic rays usually produces a coloured ray. For instance, a mixture of red and green light produces yellow light and that of green and violet, blue (see the colour insert). This means that there is a colour to correspond to each monochromatic ray (see Section 35-1), but not necessarily a monochromatic ray for each colour.

Experience shows that by mixing in definite proportions rays of the three *primary* colours (red, green and violet) one can obtain light of any colour (see the colour insert). It is interesting to note that sometimes mixing rays of two colours produces light of white colour. Such colours are termed *complementary*. An example of a complementary pair of colours is the pair of rays yellow and blue. Obviously, the mixing of rays of two colours which together contain all the colours of a rainbow will always produce white. Therefore such colours are always complementary.

### 37-4 The Colour of Objects

The colour of a body that is itself a source of light is determined by its composition, structure, the surroundings and by processes taking place in the body.

Since a body owes its colour to the radiation propagating from it, one can obtain a lot of important information about it by studying the particulars of its spectrum. The colouring of secondary light sources depends in addition on the composition of the radiation falling on them.

We recall that the *colouring of a transparent body* depends on the composition of the radiation which passes through it. When one illuminates various transparent bodies with white light, one is bound to notice that some of them remain colourless and the others appear coloured. Analyzing the spectrum of transmitted radiation with the aid of a prism, one sees that the spectrum of a colourless body contains the rays of all the colours present in the rainbow, while the spectra of coloured bodies consist of more or less wide bands of different colours, although sometimes only of a narrow band of almost a single colour.



The latter is the case of *light filters*—coloured glasses transmitting light, each of a single colour. This means that many transparent bodies absorb radiation of different colours in a different way. For instance, a red filter absorbs radiation of all colours except red, while a yellow filter absorbs only red and violet rays.

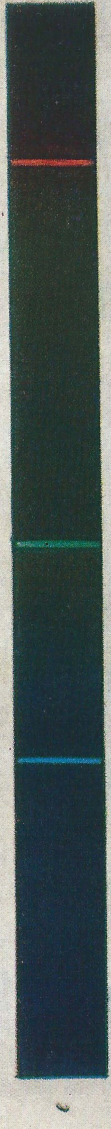
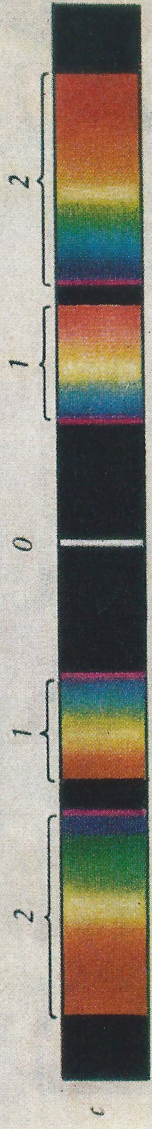
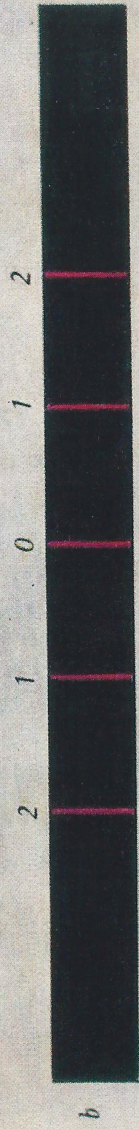
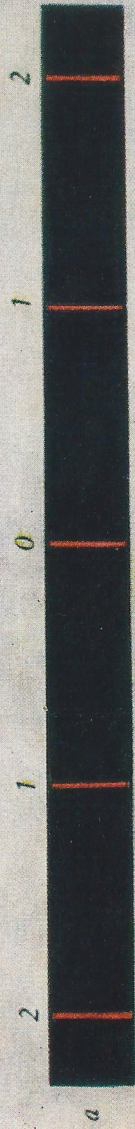
Every material has its own absorption spectrum. If a transparent material absorbs rays of all colours uniformly, it will appear colourless in transmitted light when illuminated with white light, and when illuminated with coloured light it will have the colour of that light. A body intensely absorbing light of all colours appears black to us. A body exhibiting selective absorption when illuminated with the light it transmits has the colour of that light. When the colour of the illuminating light is changed to that which the body absorbs, it turns black (i.e. becomes nontransparent).

The *colouring of a nontransparent body* in scattered light is determined by the colours of the light scattered by it. If a body scatters light of all colours of the rainbow uniformly, it appears white when illuminated with white light and of the colour of the illuminating light when the latter is coloured.

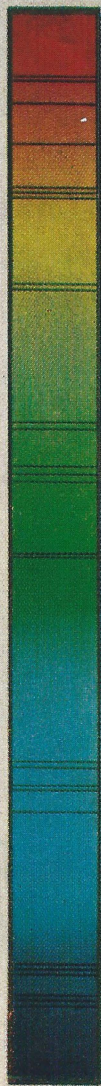
Many nontransparent bodies exhibit preferential absorption of some parts of the visible spectrum. Because of that they appear coloured when illuminated with white light. When such bodies are illuminated with light which they absorb, they appear black. Often the colour of a body is due to the colouring of its surface. The mixing of paints produces a colour different from that obtained by mixing light of the same colours. Recall that mixing yellow and blue light produces white light, but mixing yellow and blue paint produces green paint (see the colour insert). The explanation is that the green paint scatters only yellow and green light, and blue paint scatters blue and green. Therefore those two paints together scatter only green light.

Actually, by mixing three paints (yellow, blue and purple) one can produce a paint of any colour. For this reason the principal paints used in colour printing are the yellow, the blue and the purple.

It follows from the above that the colouring of a body in transmitted and in scattered light can be quite different. Since the colouring of bodies depends to a great extent on the composition of incident radiation, it is advisable to buy coloured goods, for instance fabric, in daylight.



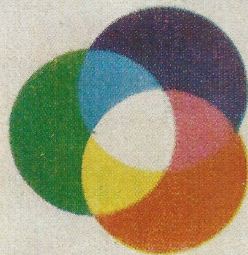
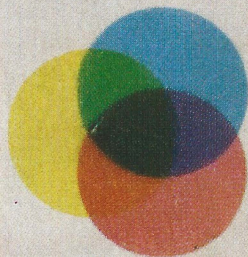




f



s



h

### 37-5 Ultraviolet and Infrared Spectra

The intensity of a spectrum can be raised with the use of lenses. The left-hand lens in Fig. 37.6 collects the rays coming from the light source and the right-hand lens assembles all the rays of one colour at a definite point on the screen.

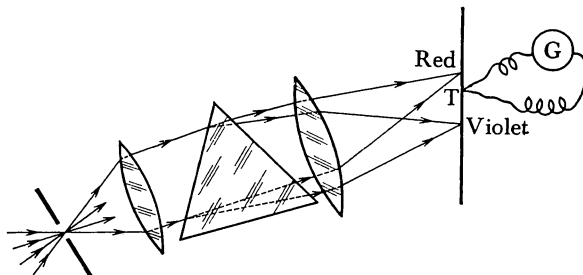


Fig. 37.6 Measuring spectral energy distribution of white light with thermocouple.

To find what rays transport more, or less, energy to the screen we use a thermocouple *T* with a soot-coated junction. The junction absorbs incident radiation and becomes hot. The resulting emf is measured with the aid of a galvanometer. The greater the energy transported by the radiation the greater the emf produced in the thermocouple.

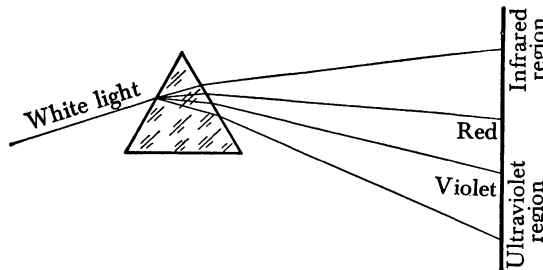
Studies of the spectrum of white light produced the result that the thermocouple junction is heated even if placed beyond the red part of the spectrum displayed on the screen. Glass absorbs the far red rays intensely and because of that lenses and prisms made of rock salt, which is transparent to red light, are used in research in the long-wave part of the spectrum. In this case the thermocouple junction becomes quite hot even when it is far away from the visible red part of the spectrum, far from where the eye is unable to see anything. This means that the red rays in the spectrum of white light are followed by invisible rays of greater wavelength than the red.

The term for these invisible rays which occupy a place beyond the red rays in the spectrum is *infrared* (from the Latin *infra* for below). They have a pronounced thermal effect and for this reason they are often termed *thermal* rays. The infrared rays experience smaller refraction than the red (Fig. 37.7) and their wavelengths lie in the range from 0.76 to 350  $\mu\text{m}$ .

It has been demonstrated in experiments that glass absorbs light strongly in the short-wave part of the spectrum as well. Accordingly, lenses and prisms made of quartz transparent to such radiation came into use for re-

search purposes. It was established that the short waves exercise a pronounced chemical effect, for instance, blacken light-sensitive paper. Such paper was found to blacken even when placed beyond the far violet part of the spectrum invisible to the human eye. The invisible rays beyond the extreme violet part of the spectrum are termed *ultraviolet*

**Fig. 37.7** There are invisible infrared rays past red rays in the spectrum of white light and ultraviolet rays past violet.



(from the Latin *ultra* for beyond). They experience greater refraction than violet rays (see Fig. 37.7), have shorter wavelengths lying in the range from 0.4 to 0.005  $\mu\text{m}$ , and exercise a pronounced chemical effect.

### 37-6 Ultraviolet and Infrared Radiation in Nature and Technology

All existing bodies emit infrared rays, since they are the result of random motion of the molecules and atoms making up the body. An increase in temperature raises the energy of infrared radiation emitted by a body.

Nearby bodies exchange radiation, with each body simultaneously emitting its own radiation and absorbing the radiation of the other bodies. The body whose temperature is the highest receives less energy in the form of radiation, than it gives up. Therefore its temperature drops. Conversely, the body with the lowest temperature, in absorbing radiation, gains more energy than it radiates itself, and so becomes hotter. Thus, all existing bodies exchange energy, this exchange tending to level out their temperatures.

Radiation transports solar energy to the Earth. Experiments have demonstrated that solar radiation includes a high percentage of energy in the form of infrared and ultraviolet rays. The energy of solar radiation is responsible for the difference in temperatures at different places on the surface of the Earth.

Infrared radiation of the Earth carries energy away from its surface, thereby cooling it. This is the cause of a drastic nightly fall in temperature in the deserts, where the atmosphere is transparent, even though the daytime temperatures are very high. In the presence of clouds the infrared radiation from the Earth is reflected back from them and the loss of energy into space is reduced. Therefore dense clouds above the ground in winter are accompanied by a rise in temperature.

Ultraviolet light present in solar radiation is strongly absorbed by the atmosphere and its energy at the surface of the Earth is comparatively low. High up in the mountains the ultraviolet component of solar radiation is much more intense.

Ultraviolet light kills bacteria, which implies that it is a good disinfectant. In small doses it is beneficial to man, tanning his skin.

In industry infrared light is used to dry materials, for instance food products, for communications in poor visibility, for taking photographs in the dark, etc. In military applications such rays are used to aim projectiles and missiles, to locate a camouflaged enemy, etc. In science infrared rays allow us to find the difference between the temperatures of individual areas on the surface of planets, for instance on Mars, to reveal the particulars of molecular structures of different substances, etc.

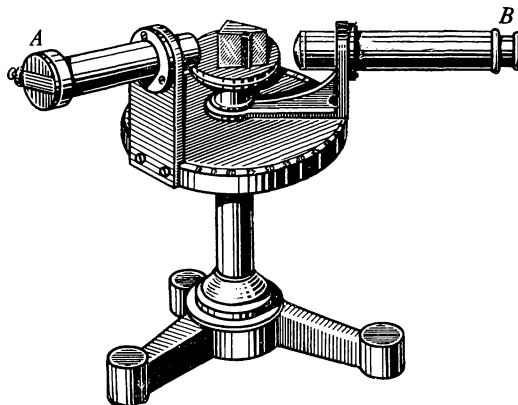
Ultraviolet light is used in photography to detect concealed lettering or erased text because many substances absorbing ultraviolet rays emit visible light. The same phenomenon is utilized in fluorescent lamps and in many other cases. Ultraviolet rays are also used to study the structure of the outer electron shells of atoms. In medicine they are used in treating certain maladies.

### 37-7 Spectroscope and Spectrograph

The instrument used to observe spectra is the *spectroscope*. The most common type, the *prismatic spectroscope*, consists of two tubes with a triangular prism placed between them (Fig. 37.8). The tube *A*, called the *collimator*, has a narrow slit whose width can be adjusted by turning a screw. A light source whose spectrum is to be investigated is placed in front of the slit. The slit is arranged in the focal plane of the collimator's lens and because of that the light rays emerge from the collimator in the form of a parallel beam. Passing through the prism the rays enter

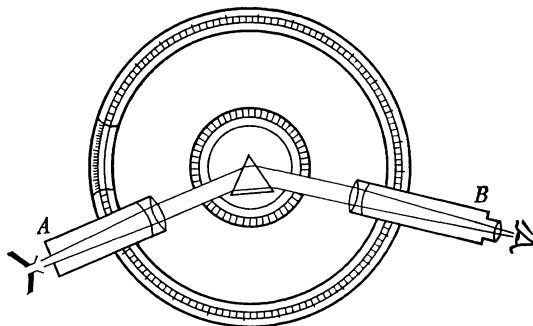
the tube *B* (see Figs. 37.8 and 37.9), through which the spectrum is observed. If the spectroscope is intended for measurements, an image of a scale with divisions is placed on

**Fig. 37.8** Spectroscope: *A*, collimator; *B*, eye-piece.

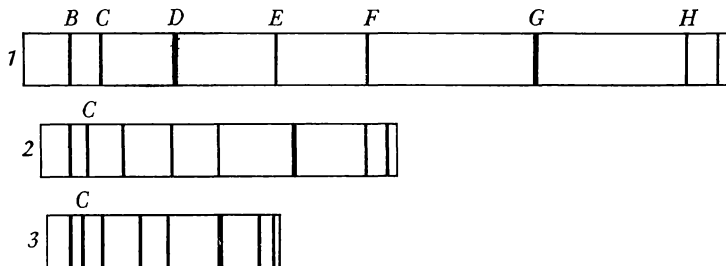


top of the image of the spectrum with the aid of a special device, thus making possible an accurate location of the coloured lines in the spectrum.

**Fig. 37.9** Schematic diagram of spectroscope.



**Fig. 37.10** Spectra of different substances (nature of dark lines is explained in Section 37-12).



The material used for the fabrication of the prism should have great dispersion (i.e. it should produce a wide spectrum). Figure 37.10 shows spectra produced with the aid of water (3), normal glass (crown) (2) and glass containing lead (flint) (1). The origin of dark lines will be explained in Section 37-12. It is evident from the figure that the best material for obtaining visible spectrum is flint glass.

When studying spectra it is often advisable first to photograph them and then to study them under a microscope.

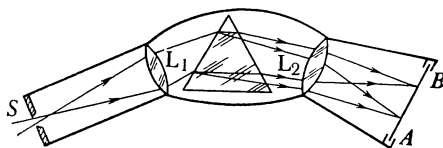


Fig. 37.11 Schematic diagram of spectrograph.

The instrument for photographing spectra is called a *spectrograph*. A schematic diagram of a spectrograph is shown in Fig. 37.11. The radiation spectrum is focused by a lens  $L_2$  on the mat glass  $AB$  which when taking photographs is replaced by a photographic plate.

### 37-8 Types of Spectra

Spectra obtained from self-luminous bodies are termed *emission spectra*. Direct observations and photographs of spectra show that emission spectra are of three types: continuous, line and band spectra.

*Continuous spectra* (see Plate (d) of the colour insert) are obtained from heated luminous solid and liquid bodies.

*Line spectra* (see Plate (e) of the colour insert) consist of narrow lines of different colour with dark intervals in between. Such spectra are often obtained from luminous gases or vapours. Luminosity can be excited in a gas by passing a current through it. To study the spectrum of a gas a glass tube containing the gas is placed in front of the spectroscope's slit and an electric current is made to flow through the gas.

Line spectra of vapours and gases can also be obtained by heating them, for instance in the flame of a burner. The same method can be used to obtain line spectra of substances which in normal conditions are in a solid or liquid state. To this end solid particles or asbestos soaked with the liquid are introduced into the flame of a Bunsen burner. The substances evaporating in the flame produce a line spectrum.



Sometimes such substances are placed into an electric arc; its red-hot electrodes are covered with a diaphragm and bright lines against the background of the weaker continuous spectrum of the arc itself are observed in the spectroscope. Note that a term frequently used for the luminous spectral lines is *emission lines*.

It was established as a result of studies of line spectra of various substances that each chemical element produces its own line spectrum different from the spectra of other elements. The line spectra of chemical elements are distinguished by their colour and by the relative position and number of luminous lines. Lines characteristic of each individual element are obtained not only in the visible part of the spectrum but also in its ultraviolet and infrared parts. The first to carry out research into line spectra were the German scientists Gustav R. Kirchhoff (1824-1887) and Robert W. E. Bunsen (1811-1899) in 1854-59.

Line spectra are produced by individual atoms of chemical elements not bounded into molecules. This radiation

Fig. 37.12 Spectrum of iodine vapour.



is the outcome of processes taking place inside the atom. The study of line spectra helped to reveal the structure of electron shells of the atoms of various chemical elements.

*Band spectra* consist of several bright bands interspaced with dark intervals (Fig. 37.12 and Plate (g) of the colour insert). Band spectra are the product of molecular radiation. In a high-resolution spectroscope the bands are resolved into several lines.

### 37-9 Absorption of Light in Gases and Vapours

It was revealed above that transparent substances absorb some of the incident radiation and that spectra of white light transmitted through such substances lack several colours, with dark lines appearing in their place. The term for such a spectrum is *absorption spectrum*.

Of great interest to us are the absorption spectra of monatomic gases having emission spectra of the line type. What rays will such a gas absorb if white light is transmitted through it?

Kirchhoff, in 1854, was the first to carry out such research. He introduced a source of sodium vapour (metallic sodium in a small crucible) or asbestos soaked with cooking salt solution into the flame of a Bunsen burner. This coloured the flame of the burner in a characteristic yellow colour corresponding to the emission of sodium vapour. Two closely spaced bright yellow lines appeared in the spectrum

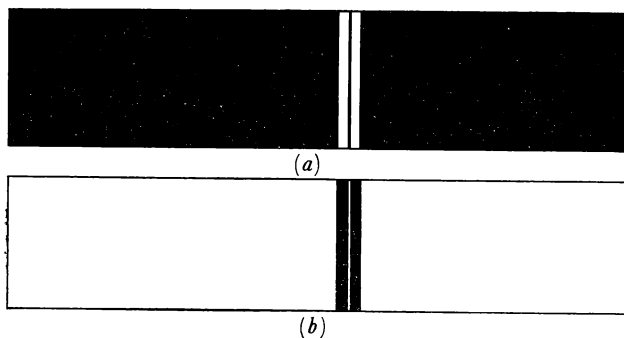


Fig. 37 13 Schematic representation of spectra of sodium vapour: (a) emission spectrum; (b) absorption spectrum.

of this radiation (Fig. 37. 13a). Next an arc torch was placed in front of the burner so that light from the arc could reach the spectroscope's slit only via the flame of the burner. In this case two dark lines appeared in the spectrum of white light coming from the arc (Fig. 37.13b) in exactly the same place as the yellow lines of the sodium emission spectrum had been.

The explanation for the appearance of these lines is that the sodium atoms out of all rays that are transmitted absorb only those that they can themselves emit. The cause of such selective absorption of radiation will be discussed in Section 38-16. While absorbing yellow rays out of the arc's radiation the sodium atoms, of course, continue to emit such rays themselves (to be more precise, atoms in the ground state absorb radiation and excited atoms emit it). But the temperature of the arc is substantially higher than that of the burner's flame and its spectrum is much brighter; so against its background the yellow lines of the sodium vapour appear dark. The spectrum still has yellow emission lines; if the electric arc is switched off, the sodium vapour spectrum will be clearly seen on the screen in place of the dark lines.

Such a phenomenon of the reversibility of spectral lines is observed in the emission and absorption spectra of many

other elements. It is expressed by Kirchhoff's law (see Section 37-10): every substance preferentially absorbs the rays it can itself emit.

### 37-10 Kirchhoff's Law of Radiation

All heated bodies—solid, liquid, or gaseous—emit radiation. The radiation is the result of atomic and molecular excitation in the course of their random thermal motion, that is, the energy of this radiation is liberated at the expense of the internal energy of the body. Radiation completely determined by the temperature of the body is termed *thermal radiation*.

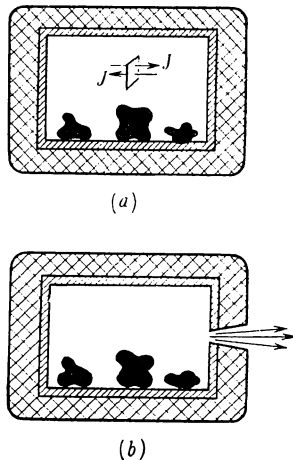
The properties of thermal radiation of various bodies are determined by their temperature and depend on their nature. The radiation of different bodies at an identical temperature is not the same. For example, a metal rod in the flame of a Bunsen burner appears brighter than a quartz rod, the luminosity of the flame itself being quite weak. Thermal radiation of a body is determined by its *radiant emittance* (denoted by the letter  $e$ ). The measure of emittance is the energy of radiation emitted by a unit area of the body's surface per unit time.

All bodies can absorb incident radiation. The energy of absorbed radiation is transformed into the internal energy of the body. It is an experimental fact that some bodies are good absorbers and others poor. Accordingly, every body is characterized by its *absorptance* (denoted by the letter  $a$ ). Absorptance is expressed by the ratio of radiation absorbed by the body to the incident radiation. Absorptance depends on the nature of the body, on the state of its surface, as well as on the radiation wavelength. We recall that a body is said to be *black* if it completely absorbs all incident radiation. The absorptance of black body is unity; for other bodies,  $a < 1$ ; for an ideal mirror  $a = 0$ . The material with characteristics close to those of a black body in the visible part of the spectrum is soot.

Let us find the relation between the absorptance and the radiant emittance of a body. Suppose various heated bodies are placed inside an enclosed space with good thermal insulation against outside influence (Fig. 37.14a) and that the bodies can exchange energy by means of radiation.

If initially the bodies had different temperatures, the hotter ones will radiate more energy than they absorb and will cool down as the result. The colder ones will get hotter. After some time the temperature of all the bodies and of

Fig. 37.14 (a) Deducing Kirchhoff's thermal radiation law: (b) opening in uniformly heated cavity is model of black body.



the walls bounding the space will level out and a thermodynamic equilibrium will be established between the bodies and the radiation in which all the bodies, including the walls, radiate as much energy as they absorb. The space inside will be uniformly filled with electromagnetic radiation of various wavelengths and intensities, the waves moving at random in all directions just like gas molecules moving in a closed space. There will be no directional energy transport in the system—all the directions will become equivalent. Any arbitrary area of  $1 \text{ m}^2$  (see Fig. 37.14a) inside the space will receive in the period of 1 s equal amounts of radiation energy  $J$ , the same energy falling on the opposite side of the area.

This energy  $J$  will fall in the period of 1 s on every  $1 \text{ m}^2$  of the surface of any body belonging to the system. Each body will absorb a part of this energy depending on its absorptance:  $a_1 J$ ,  $a_2 J$ , etc. In a state of equilibrium each body radiates from  $1 \text{ m}^2$  of its surface as much energy as it absorbs. Denoting the radiant emittances of the bodies by  $e_1$ ,  $e_2$ , etc., we obtain  $e_1 = a_1 J$ ,  $e_2 = a_2 J$ , etc. Hence

$$e_1/a_1 = e_2/a_2 = \dots = J$$

Let one of the bodies be black. Then for it  $e_b/a_b = J$ , and  $e_b = J$  since  $a_b = 1$ . Therefore

$$e_1/a_1 = e_2/a_2 = \dots = e_b \quad (37.1)$$

This important relation is the expression for *Kirchhoff's radiation law* deduced by him from theoretical considerations in 1860: the ratio of the radiant emittance of an arbitrary body to its absorptance at a specified temperature is independent of the nature of the body and is equal to the radiant emittance of a black body.

Hence, the greater the absorptance of a body the greater its radiant emittance. At a specified temperature a black body has the greatest radiant emittance. Accordingly, at an equal temperature a black body emits more light.

There is yet another important conclusion to be drawn from the aforesaid; in equilibrium a unit area of the surface of any body radiates in a unit of time the same energy as that of a black body. Indeed, out of the incident radiation  $J$  a part  $aJ$  is absorbed by the body and the rest  $(1 - a)J$  is reflected or transmitted together with the energy  $e$  radiated by the body equal to the absorbed energy  $aJ$ :

$$(1 - a)J + e = (1 - a)J + aJ = e_b$$

Thus, the radiation from anywhere inside such a space is the same as the radiation of a black body. Accordingly, if we make a hole in the enclosure small enough to preclude noticeable deviation from thermal equilibrium (Fig. 37.14*b*), this hole will have emission characteristics of a black body. Hence, a hole in a uniformly heated cavity serves as a good model of a black body. It can also be easily seen that such a hole absorbs all incidental outside radiation: a ray of light which enters the cavity through the hole experiences multiple reflections and is eventually completely absorbed in it (indeed, a small hole in a closed box appears even blacker than soot).

The radiation of bodies in a closed oven is close to equilibrium. Accordingly, if one peers into the oven through a small hole, one sees various bodies made, for instance of graphite, of metal, of quartz, as equally bright, so that they can be barely distinguished from one another. However, should these bodies be withdrawn from the oven, their emission would become nonequilibrium and the luminosity of each body would depend on its radiant emittance. Despite the fact that the initial temperature of all bodies is the same, the part made of graphite shines brighter.

Obviously, bodies with greater absorptance are heated by radiation more rapidly, but at the same time they lose heat more quickly, because they emit more than bodies of low absorptance. (Explain why the surface of a vacuum flask is made with a mirror finish.)

Kirchhoff's law is valid not only for the total energy of radiation at all wavelengths but also for each waveband. Indeed, all the arguments used in deducing Kirchhoff's law can be repeated in an imaginary case where the cavity of Fig. 37.14*a* is divided by numerous light filters, each transmitting radiation of a specified waveband. As a result we obtain for the radiant emittance and absorptance inside this band a relation similar to (37.1):

$$e_{\lambda_1}/a_{\lambda_1} = e_{\lambda_2}/a_{\lambda_2} = \dots = e_{\lambda_b} \quad (37.1a)$$

Hence, if some body strongly absorbs radiation in some part of the spectrum, it must also strongly emit radiation in the same part of the spectrum. Accordingly, there is another expression for Kirchhoff's law: every body preferentially absorbs the rays it can emit itself (see Section 37-9), so that the locations of the corresponding lines in the absorption and the emission spectra coincide.

### 37-11 The Stefan-Boltzmann, Wien and Planck Radiation Laws

The ideal black body can conveniently be used as a standard emitter (in practice black platinum or a hole in a uniformly heated cavity is used), because its emittance is fully determined by its temperature. The *Stefan-Boltzmann law* establishes this dependence: the radiant emittance of a black body is directly proportional to the fourth power of its temperature:

$$e_b = \sigma T^4 \quad (37.2)$$

where  $\sigma = 5.67 \times 10^{-8} \text{ J}/(\text{m}^2 \cdot \text{s} \cdot \text{K}^4)$  is the Stefan-Boltzmann constant.

Efforts have been made to calculate theoretically the frequency distribution of thermal radiation using both the methods of thermodynamics and electrodynamics. According to the theory of electromagnetic waves the spectral density of energy radiated by a heated body should increase with frequency. This means, for example, that the radiation of an incandescent lamp should contain a large proportion of ultraviolet rays.

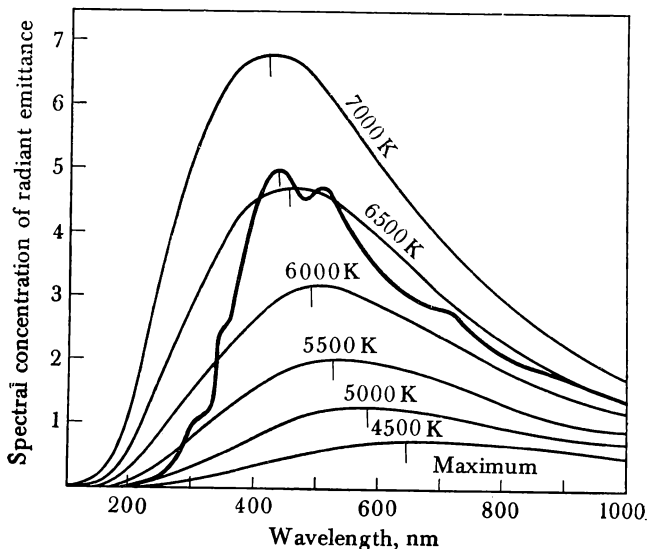
However, it was established as the result of experiments that, although the spectral density of the energy of radiation does indeed rise at first with the frequency, it then peaks and begins to decline, tending to zero at very high frequencies. The scientist who succeeded in bringing the theory in line with experiment was Planck, who in 1900 introduced the concept of the quantum properties of radiation. He made the assumption that a substance emits energy not continuously but in definite portions (quanta) of energy proportional to the radiation frequency. The magnitude of energy quanta is determined from Planck's formula:  $\epsilon = hf$  (see Section 31-3), where  $f$  is the frequency and  $h$  is Planck's constant. The higher the frequency the greater the quantum of radiated energy.

At low temperatures the energy of thermal motion of the particles making up a body is not enough to generate quanta of high energy. The higher the temperature of a body the greater the probability of the generation of high-energy quanta in the radiation of the body, the greater the intensity and the diversity of the radiation, and the higher the frequencies (shorter the waves) covered by its spectrum.

Figure 37.15 shows the wavelength distribution of the radiation of a black body obtained by Planck (Planck's curves) for several temperatures. It is evident from the

figure that the wavelength corresponding to maximum spectral density becomes shorter the higher the temperature. This relationship was obtained by the German physicist Wilhelm Wien (1864-1928) in 1893 and is known as the

**Fig. 37.15** Energy distribution in the spectrum of black body at different temperatures and energy distribution in solar spectrum (solid curve).



*Wien law:* the product of the wavelength corresponding to the maximum spectral density of radiation of black body and of its absolute temperature is a constant:

$$\lambda_{\max} T = b \quad (37.3)$$

where  $b = 0.002\,898 \text{ m} \cdot \text{K}$  is Wien's constant.

The spectral distribution of the energy of the radiation of some real body can be appreciably different from the one shown in Fig. 37.15 for the black body at the same temperature, but the general tendency is usually the same. If one finds from experiment the wavelength corresponding to maximum spectral density in the radiation spectrum of a body, it is then possible to estimate its temperature. This method can be used to find the temperature of, for instance molten metal, of the filament of an incandescent lamp, etc. The term for this method of measuring the temperature of a source of radiation is *optical pyrometry*.

Planck's theory of thermal radiation is in good agreement with experimental data: specifically, the laws of Stefan-Boltzmann and Wien are particular cases of Planck's law. Here we have the first proof of the quantum properties of radiation.

### 37-12 Solar and Stellar Spectra

Kirchhoff's law explained the origin of dark lines in the spectrum of solar radiation (see Plate (f) of the colour insert). They were discovered and described in 1817 by the Bavarian instrument maker Joseph Fraunhofer (1787-1826), who was the first to use a diffraction grating to obtain a spectrum.

The dark lines, which were named *Fraunhofer lines*, occupy strictly defined places in the solar radiation spectrum. The more pronounced lines were denoted by the letters *A*, *B*, *C*, *D*, etc. (see Fig. 37.10). It was established that Fraunhofer lines are absorption lines of the vapours and gases in the external layers of the photosphere which are colder than its internal layers. (Compare with Kirchhoff's experiment.) Thus, for instance, the line *D* is the yellow line of sodium vapour, the lines *C* and *F* are absorption lines of hydrogen, etc. Similar absorption lines were also discovered in stellar spectra.

As was revealed above, practically all the visible radiation of the Sun is emitted by the photosphere. The radiation of the deeper lying layers is absorbed in the photosphere and does not pass through it. The external layers of the Sun's atmosphere (the chromosphere and the corona), although they are hotter than the photosphere, are too rarefied to contribute substantially to the visible radiation of the Sun (the luminance of the chromosphere is hundreds of times and of the corona a million times less than that of the photosphere).

The main part of the photosphere's radiation is emitted by its internal hotter layers. They are well insulated thermally from outer space by external layers and because of that their radiation is close to equilibrium. Accordingly, the Sun's radiation should approach that of a black body.

Figure 37.15 depicts Planck's curves for several temperatures and the real spectral distribution of energy in the solar spectrum; it is seen to correspond to the radiation of a black body having a temperature of 6000-6500 K. The solar radiation is most intense in the blue-green part of the spectrum in the wavelength interval from 430 to 500 nm.

To sum up, the laws of thermal radiation can be used to estimate the temperature of the Sun and of other stars. Their temperature can be found with the aid of the Stefan-Boltzmann law if the area of the emission surface and the total radiated energy are known. Another possibility is to select an appropriate Planck's curve or to use the Wien



law, to find the temperature from the wavelength corresponding to the maximum spectral density of radiation. The temperature of the photosphere of the Sun measured with the aid of different methods is close to 6000 K.

Stars have different temperatures. One can establish this fact simply by their colour. Look at the night sky strewn with bluish, white, yellow and red stars. Obviously, the hottest among them are the light-blue ones (their temperature exceeds 30 000 K) and the coolest are the red ones (around 3000 K). The Sun is a yellow star. Note that with the naked eye one is able to distinguish the colours only of the brightest stars.

Other signs of the difference in the temperature of individual stars are the difference in intensities and in the number of lines belonging to specific chemical elements. Thus, the spectra of very hot stars contain bright emission lines of helium and nitrogen, and the spectra of cooler stars strong absorption lines of various molecular compounds.

The solar spectrum extends far into the shortwave and the longwave parts of the spectrum. In the shortwave part the intensity of the continuous spectrum falls off rapidly, the dark Fraunhofer lines being replaced by bright emission lines, of which there are several thousands.

The intensity of the solar spectrum in the longwave part decreases more slowly with the frequency than that of a black body with a temperature of about 6000 K, and in the radio-frequency band the Sun radiates like a black body at a temperature of  $10^6$  K. In contrast to visible radiation solar radiofrequency radiation can change greatly in intensity. For instance, solar flare is accompanied by a flash of radiofrequency radiation—a great (sometimes by millions) rise in the power of radiation at certain radiofrequencies.

### 37-13 Spectroscopic Analysis

Every chemical element has its own characteristic emission spectrum. This makes it possible to identify the chemical elements making up a body by the line spectrum of its vapours. The term for such a method determining the chemical composition of a substance is *qualitative spectroscopic analysis*.

Spectroscopic analysis is widely used in science and industry. This is one of the quickest and easiest methods of determining the composition of various chemical compounds. It boasts extremely high sensitivity and enables the pres-

ence of minute quantities of chemical elements to be established. The amounts of the substance analyzed needed for spectroscopic analysis is also very small (often  $10^{-8}$ - $10^{-9}$  g is enough).

The use of spectroscopic analysis makes it possible to find the composition of gases and vapours no matter how distant they are, provided their rays are able to reach the spectral instrument. For this reason this method is widely used in astronomy to establish the chemical composition of the Sun and the stars, their temperature, their motion in space (see Section 37-14), etc.

The first remarkable feat accomplished by spectroscopic analysis was the discovery of new chemical elements. The founders of spectral analysis, Bunsen and Kirchhoff, used it to discover the new alkali metals rubidium and caesium. Subsequently, other elements were also discovered, namely, indium and thallium. Especially interesting was the discovery of helium. Initially it was discovered during an analysis made of the spectrum of a solar protuberance in 1868, whence the name of the element (from the Greek *helios* for Sun). Helium lines were found in 1881 in the spectrum of gases from Mount Vesuvius, then in some minerals, and in very small quantities in the atmosphere of the Earth. Only in 1905 did scientists manage to produce small amounts of helium.

In accordance with Kirchhoff's law the spectral analysis of gases and vapours can be made with the aid of their absorption spectra. Thus, the study of the location of Fraunhofer lines in the solar spectrum led to the conclusion that the Sun is made up of the same elements as the Earth.

When carrying out spectroscopic analysis recourse is made to special tables and atlases of spectral lines which contain information on the precise location of the spectral lines of each chemical element and on wavelengths corresponding to it. In some cases a spectral analysis is made on the basis of a comparison of the spectra of the material being studied and of a standard sample with a known composition of chemical elements.

At present methods of *quantitative spectroscopic analysis* have been developed involving the determination of the presence of a chemical element in the specimen from the luminous intensity of its spectral lines.

The main advantages of spectroscopic analysis—very high sensitivity, simplicity and speed of analysis—make it a very convenient tool for use in metallurgy, in machine tool production, in chemistry and geology, in medicine and biology, and in various other areas of science and technology.

### 37-14 The Doppler Effect

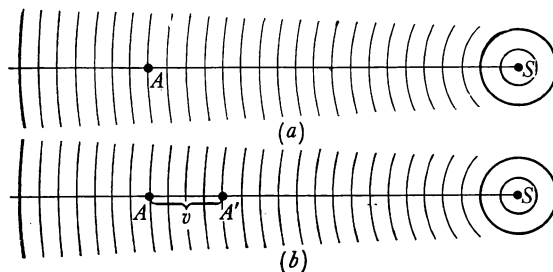
The Austrian mathematician and physicist Christian J. Doppler (1803-1853) in 1842 established that the motion of a source of oscillations in the direction of the observer increases the frequency of the oscillations received by the observer, and the motion in the direction from the observer causing a decrease in the frequency. This phenomenon, termed the *Doppler effect*, can be observed, for example, when a signalling train or a car speeds past us. When the train approaches, we hear its whistle as having a higher tone, and as it moves away a lower tone than when the train is at rest.

The term for the change in frequency of oscillations in a wave (in wavelength) registered by an observer when the source of the waves approaches or moves away from him is Doppler effect.

The Doppler effect is a feature of all waves, including electromagnetic.

Let the source  $S$  radiate electromagnetic waves (in a vacuum) with the frequency  $f_0 = c/\lambda_0$ , where  $c$  is the velocity of light and  $\lambda_0$  is the wavelength. Over one second the waves propagating from the source  $S$  travel a distance equal

**Fig. 37.16** (a) Observer at point  $A$  is at rest with respect to wave source  $S$ ; (b) observer moves in direction of source at speed  $v$ ; in one second he moves from point  $A$  to point  $A'$ .



to  $c$ , and  $c/\lambda_0$  waves pass through a point  $A$  (Fig. 37.16a) in that time. Accordingly, an observer at rest at this point will register waves with a frequency  $f_{\text{obs}} = c/\lambda_0 = f_0$ .

Now let the observer move towards the source at a speed  $v$ . Now in one second he will approach the source by a distance  $AA'$  numerically equal to  $v$  (Fig. 37.16b), so  $v/\lambda_0$  more waves will move past him per second than when he was at rest. Accordingly, the frequency of the waves registered by the observer will be greater than  $f_0$  by  $v/\lambda_0$ :

$$f_{\text{obs}} - f_0 = \Delta f = v/\lambda_0 \quad (37.4)$$

It can easily be inferred that, if the observer moves away from the source at a speed  $v$ , the frequency registered by the observer will be by  $v/\lambda_0$  less than  $f_0$

$$f_0 - f_{\text{obs}} = \Delta f = v/\lambda_0 \quad (37.5)$$

Since  $\lambda_0 = c/f_0$  it follows that  $\Delta f = f_0 v/c$ , or

$$\Delta f/f_0 = v/c \quad (37.6)$$

There is a definite wavelength of a wave in a medium to correspond to a definite frequency; therefore the Doppler effect can be regarded as the result of a change in wavelength registered by the observer. Using the relations  $f_{\text{obs}} = c/\lambda_{\text{obs}}$  and  $f_0 = c/\lambda_0$ , we can obtain from (37.4) or from (37.5)

$$\Delta\lambda/\lambda_0 = v/c \quad (37.7)$$

The wavelength  $\lambda_{\text{obs}}$  will be less than  $\lambda_0$  by  $\Delta\lambda$  if the observer approaches the source (as he moves against the waves they appear somewhat compressed to him), and  $\lambda_{\text{obs}}$  will be greater than  $\lambda_0$  by  $\Delta\lambda$  if the observer moves away from the source (in this case they appear somewhat rarefied to him). Knowing the change in the wavelength  $\Delta\lambda$  (or in the frequency  $\Delta f$ ) registered by the observer, it is possible to determine the speed at which the observer and the source of the waves move towards, or away from, each other.

It should be noted that formulae (37.4)-(37.7) are valid only for speeds of motion  $v$  much less than the velocity of light  $c$ . A more rigorous derivation of the formula for the Doppler shift involves the theory of relativity (see Chapter 39).

It was established in the course of studying stellar spectra that the stars move with respect to the solar system, since the lines of known elements in their spectra are displaced as compared with their location in the spectrum of a laboratory (stationary) source of radiation. The Doppler shift of the spectral lines is used to estimate the speed of approach or recession of the stars.

When spectra of distant star systems, galaxies, were obtained, a displacement of the spectral lines in the direction of longer wavelengths (towards the red part of the spectrum) was discovered. This is called the *red shift*. The magnitude of the red shift, together with the recession speed corresponding to it, were found to increase with the distance to the galaxy. This has led to the assumption that the galaxies are moving away from each other.

### 37-15 X Rays and Their Practical Uses

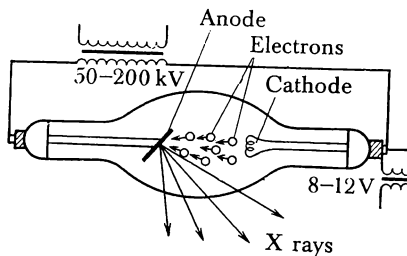
In 1895 the German physicist Wilhelm Konrad Roentgen (1845-1923) discovered that mysterious rays were produced in the tube used for the cathode-ray production. These rays penetrate through glass, air and through many other bodies nontransparent to visible light. Those rays were subsequently termed *X rays*.

The X rays themselves are invisible, but they excite the luminescence of many substances and act intensely on photosensitive materials. Accordingly, for their study special screens are used which light up when X rays fall on them. This property enabled Roentgen to discover them.

X rays are produced when fast electrons decelerate. There is a magnetic field surrounding flying electrons because the motion of an electron is an electric current. When an electron colliding with an obstacle brakes sharply, its magnetic field changes rapidly and an electromagnetic wave is radiated into space whose wavelength is the shorter the greater was the speed of the electron before its collision with the obstacle.

X rays are produced in special two-electrode tubes (Fig. 37.17) to which high voltage of the order of 50 000-200 000 V

Fig. 37.17 Schematic representation of X-ray tube.



is applied. The electrons emitted by the glowing cathode of the X-ray tube are accelerated in a strong electric field in the space between the anode and the cathode and strike the anode at a great speed. This causes the anode surface to emit X rays which pass through the tube's glass. The radiation of an X-ray tube has a continuous spectrum.

X-ray tubes with a hot cathode themselves act as rectifiers and accordingly they can use an ac power supply.

If the electrons are accelerated in the field to speeds sufficient for them to penetrate inside an atom of the anode and knock one of the electrons out of its inner shell, the place of the latter will be taken by an electron of one of the

more remote shells, the process being accompanied by the emission of a high-energy quantum. Such radiation has a definite spectrum peculiar only to each specific chemical element and for this reason it is termed *characteristic*. Characteristic radiation has a line spectrum, which is superimposed on the continuous bremsstrahlung (from the German word for braking radiation) spectrum. With the increase in the atomic number of the element in the Mendeleev Periodic Table its X-ray spectrum is displaced in the direction of shorter waves. Light elements (notably, aluminium) have no characteristic X-ray spectrum at all.

X rays are customarily distinguished by their *hardness*: the shorter the wavelength of X rays the *harder* they are. The hardest X rays are emitted by heavy atoms.

An important property of X rays is their ability to penetrate through many substances impervious to visible light. The harder the X rays the less they are absorbed and the greater their *penetration ability*.

The absorption of X rays in a substance depends on its atomic composition: the atoms of heavy elements are effective absorbers of X rays, no matter which substances they belong to.

Like all other electromagnetic waves, X rays are deflected by neither magnetic nor electric fields.

The refractive index of X rays is very close to unity and they experience almost no refraction when crossing a boundary between two media. This property of X rays, together with their high penetration ability, is utilized for practical purposes.

If a body is placed between an X-ray source and a luminescent screen, a dark shadow of the body will appear on the screen. If there is a void inside an otherwise homogeneous body, the corresponding place on the screen will show up as a brighter spot. This phenomena is used to detect internal defects in products (defectoscopy).

When a body of a nonhomogeneous molecular composition is illuminated with X rays, different parts of it absorb different amounts of X-ray radiation and we are able to see the contours of those parts on the screen. Thus, X-raying a hand, we clearly see a dark image of its bones on the screen (Fig. 37.18).

It is often more convenient to take X-ray photographs than to use a luminescent screen. To this end the body under investigation is placed between an X-ray tube and a closed film holder containing photographic film, and for a short time X rays are passed through it. After exposition the film is developed in the usual way.

Fig. 37.18 X-ray, photograph of hand.

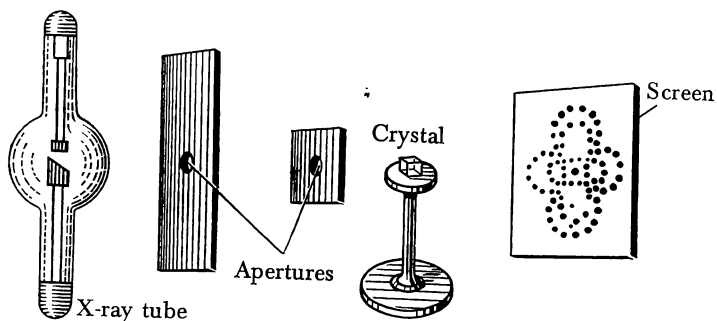


X rays are widely used in medicine: in the diagnosis of various maladies (tuberculosis, for instance), in assessing the nature of bone fractures, for detecting foreign objects in the body (for instance, a bullet lodged in the body), etc.

X rays adversely affect the evolution of cells. This property is used in the treatment of malignant tumours. However, the same property makes a prolonged or a too intensive treatment, especially with hard X rays, the cause of dangerous illnesses.

For an appreciable time after X rays were discovered all efforts to establish their wave properties—to observe their diffraction and to measure their wavelength—failed. All attempts to use diffraction gratings designed for measuring wavelengths of light waves proved unsuccessful. In 1912 the German physicist Max von Laue (1879-1960) suggested the use of natural crystal lattices for the diffraction of X rays. It was demonstrated in experiments that a narrow beam of X rays, after passing through a crystal, produces an intricate diffraction pattern consisting of a group of spots on the screen or film (Fig. 37.19).

**Fig. 37.19** X-ray diffraction.



Studies of the diffraction pattern obtained with rock salt crystals enabled the wavelength of X rays to be determined, since the distance between the sites of this lattice was known beforehand. The wavelength of X rays used in that experiment proved to have been several tenths of a nanometre. Further studies produced the result that the wavelengths of X rays lie in the range from 100 to 0.1 angstroms ( $1 \text{ \AA} = 0.1 \text{ nm}$ ). Thus, even the wavelengths of soft X rays are shorter by two orders of magnitude than those of visible light.

Now it is obvious why diffraction gratings are useless for studying X rays: the wavelengths of X rays are too short for them and there is no diffraction. On the other hand, the lattice constants of natural crystals are comparable to the

wavelengths of X rays, that is, crystals can act as ready-made diffraction gratings for them.

Laue's experiments demonstrated that X rays are electromagnetic waves. The diffraction of X rays is used to determine their wavelengths (*X-ray spectroscopic analysis*) and vice versa: by passing X rays of known wavelengths through a crystal, one can, by analyzing the diffraction pattern, find the arrangements of atoms and the interatomic distances in the lattice (*X-ray structural analysis*).

### 37-16 The Electromagnetic Spectrum

We already know that Maxwell developed the theory of electromagnetic phenomena and presented arguments in favour of the existence of electromagnetic waves, while Hertz produced and studied such waves in experiment.

The works of Hertz, Popov, Lebedev and other scientists substantiated Maxwell's theory and proved that it is possible to produce with the aid of an oscillatory circuit electromagnetic radiation with wavelengths from several kilometres to 6 mm. One conclusion to be drawn from Maxwell's theory was that luminous radiation is in fact very short electromagnetic waves produced by natural oscillators, atoms and molecules.

Thus, by the end of the last century known electromagnetic radiation included the ranges from several kilometres to 6 mm and from 0.3 mm (infrared) to 0.01  $\mu\text{m}$  (ultraviolet). Next came X rays, which later turned out to be very short electromagnetic waves.

The study of radioactive phenomena led to the discovery of electromagnetic radiation, with wavelengths still shorter than those of X rays. This radiation was termed  $\gamma$ -radiation.

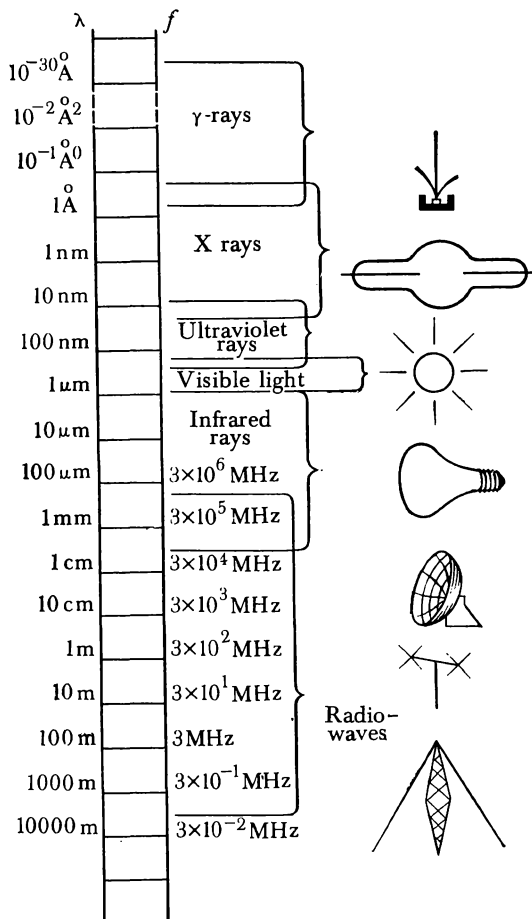
Later electromagnetic waves were obtained experimentally which filled in the originally existing gaps in the electromagnetic wave spectrum.

The scale of known electromagnetic waves is presented in Fig. 37.20. Electromagnetic waves are classified according to the type of their excitation. Overlapping in parts of the scale means that the corresponding wavelengths can be produced by two methods. For instance, the 0.1 mm waves can be produced both with the aid of an artificial oscillator and in the process of thermal radiation. Of course, the physical properties of both are absolutely identical, because they are determined by the wavelength and not by the method of excitation.



It can be seen from Fig. 37.20 that the visible range constitutes only a very small part of the electromagnetic wave spectrum.

Fig. 37.20 Electromagnetic spectrum.



Research into the properties of electromagnetic radiation is of paramount importance for adding detail to our conceptions of the structure of matter. Thus, the study of infrared, visible and ultraviolet radiation helped to find out the facts about the structure of molecules and of external atomic electron shells; the study of X-ray radiation produced information on the structure of inner atomic electron shells and on the structure of crystals, the study of  $\gamma$ -radiation continues to produce valuable information on the structure of atomic nuclei.

### 37-17 Types of Cosmic Radiation

Up to the 1940s almost all information about celestial bodies was gained by using optical methods. The reason for this is that the atmosphere of the Earth transmits only electromagnetic waves with wave length from  $0.3\text{ }\mu\text{m}$  to several micrometres and radiowaves with wavelengths from several centimetres to tens of metres. The atmosphere of the Earth is impenetrable for waves of other parts of the electromagnetic wavelength scale. At the same time electromagnetic waves of all ranges from radiowaves to  $\gamma$ -radiation are radiated in the universe.

Cosmic radiofrequency radiation was first discovered in the 1930s in the course of studies of thunderstorm interference to radio reception. In the forties and fifties the search was started for the sources of cosmic radiofrequency radiation and study in this field was intensified. Radar receivers were used to this end; subsequently the building of radiotelescopes with enormous cup-shaped antennas and sensitive radiation receivers began. Rapid progress in radioastronomy resulted in a series of important discoveries.

It was discovered that neutral cold hydrogen, which is the main component of interstellar gas but is invisible in the optic range, emits monochromatic radiofrequency radiation with a wavelength of 21 cm. This helped in determining the distribution of hydrogen in our Galaxy, including even distant parts of it hidden by dust clouds which, however, are transparent for radiowaves.

Galaxies have since been discovered with radiofrequency radiation millions of times as powerful as that of our Galaxy (they are termed *radiogalaxies*). It has been established that this strong radiofrequency radiation is not of a thermal nature. It is due to gigantic explosions involving the ejection of masses of material millions of times as great as that of the Sun. The charged particles ejected in the course of the explosion and flying at high speeds move in the interstellar magnetic field in curvilinear trajectories, that is, with acceleration. Accelerated charge motion is, as we know, accompanied by radiation of electromagnetic waves. This nonthermal radiation is termed *magnetic bremsstrahlung*, or *synchrotron radiation*, because it is observed in synchrotrons (accelerators of charged particles). The study of synchrotron radiation provides valuable information on the motion of fluxes of cosmic particles and on interstellar magnetic fields. Usually it is the radiowaves which are studied, but when the particles move at very great speeds or in sufficiently strong mag-

netic fields, they can emit visible, ultraviolet and even X-ray radiation.

A method used very extensively for registering cosmic radiation from the infrared to X-ray is the photographic method. In addition, thermocouples, thermoresistors and photoelectrical devices discussed in the following chapter are used as receivers of cosmic radiation.

As was revealed above the atmosphere of the Earth intensely absorbs shortwave radiation. Only radiation close to the ultraviolet spectrum reaches the ground and it, too, is greatly attenuated. Because of that shortwave cosmic radiation can be studied only from rockets and satellites. Such studies have made it possible to investigate the ultraviolet part of the solar spectrum and very hot stars with a temperature up to 30 000 K-powerful emitters of ultraviolet radiation.

Since the temperature of the solar corona is about  $10^6$  K (see Section 8-7) it should, in accordance with the laws of thermal radiation, be a source of X-ray radiation. The very first experiments carried out with rockets proved this to be true. It turned out that the X-ray radiation of the Sun is variable. Bursts of X-ray radiation were observed to accompany chromospheric flares. The explanation of this is that fast electrons ejected during the flash emit X rays in the process of collision with other particles of the solar atmosphere, as well as during braking in strong magnetic fields of the active regions (synchrotron radiation). Note that solar X-ray radiation is the principal ionizing agent of the upper layer of the Earth's atmosphere, the ionosphere.

From spacecraft it has been possible to detect X-ray radiation of various distant objects (galaxy nuclei, neutron stars, etc.).

## 38

# Phenomena Arising from Quantum Nature of Light

### 38-1 Wave and Quantum Properties of Radiation

This chapter deals with phenomena involving various transformations of radiation energy.

When radiation interacts with matter, in some cases its wave properties are predominant and in others its quantum properties. It follows from formula (31.1) that at low frequen-

cies the magnitude of the energy quanta is very small and it is therefore very difficult to establish the quantum nature of the interaction between radiation and matter. It should not, however, be imagined that the radiowaves possess no quantum properties at all: the emission and absorption of radio waves by atoms and molecules complies to quantum law. This has been confirmed by experiment.

Quantum properties play an essential part in the emission and absorption of X rays, whose quanta are many millions of times more energetic than the quanta of radiowaves. On the other hand, it took a long time to discover the wave properties of X rays.

When infrared, visible and ultraviolet rays interact with matter, both their wave and quantum properties come into play. For instance, reflection and refraction are described in terms of wave theory, and the explanation of the transformation of the energy of optic radiation into electric energy (the photoelectric effect) is based on quantum theory.

In short, any radiation exhibits simultaneously both wave and quantum properties. To explain any particular phenomenon one should make use of the property of radiation that exercises the greater effect on the development of the phenomenon and with the aid of which the phenomenon can be explained more easily and more clearly.

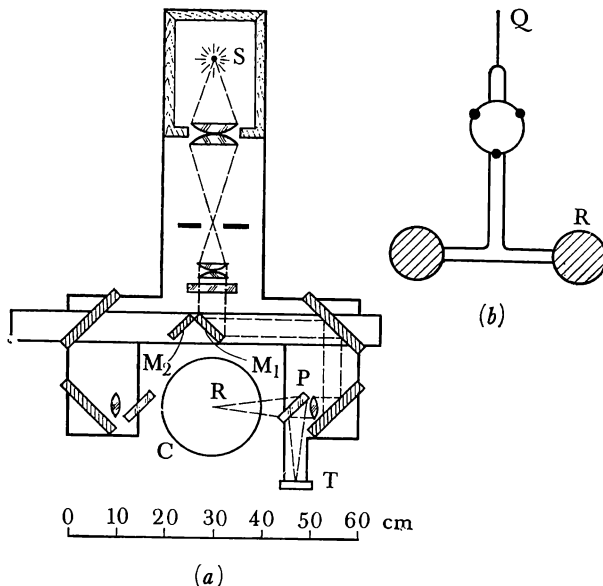
### 38-2 The Pressure of Light

It is a corollary of Maxwell's theory that light falling on a body exerts pressure on it. Accordingly, a body of a small enough mass can be set in motion by the force of light pressure. This force is, however, very small. Maxwell's calculations demonstrated that on the Earth the pressure of solar light on a square metre of a black surface normal to the light is  $4.5 \times 10^{-6}$  N. It is very difficult to detect and measure such a force in experiment, since the heating of one side of the body's surface by light increases the pressure of surrounding air on this surface by an amount many times exceeding the pressure of the light.

The first to surmount the difficulties of such an experiment was the Russian physicist P.N. Lebedev (1866-1912) in 1900. A schematic diagram of his apparatus is shown in Fig. 38.1. The light from the source S, after several reflections strikes a thin light disk R attached to a bracket. The bracket (Fig. 38.1*b*) hangs on an extremely thin quartz filament in a chamber C, in which a high vacuum is created. The force of the pressure of light on the disk is determined

from the angle of rotation of the bracket. To measure the radiation energy the plate  $P$  reflects a definite portion of the light on the thermocouple  $T$ . By moving the mirrors  $M_1$  and  $M_2$  it is possible to change the path of the rays so as to make them fall on the opposite surface of the disk.

Fig. 38.1 Lebedev's experiment.



Lebedev's experiments fully confirmed Maxwell's electromagnetic theory of light. Subsequently Lebedev measured the pressure of light on gases, which proved to be substantially less than its pressure on solids. Lebedev's experiments are an example of physical experiment of the highest precision.

For very small particles the force of the pressure of light on them can exceed the force of gravity. Observations of comets showed that as a comet approaches the Sun it often "grows" a tail, but always in the direction opposite to the Sun (Fig. 38.2). Kepler already had supposed the formation of a comet's tail to be the result of the pressure of solar light. This explanation was supported by Lebedev's experiments. (Note that the solar wind is also a factor of paramount importance in the formation of comet tails.)

Maxwell used the electromagnetic theory to explain the pressure of light on a body. However, it may be explained more easily and in a graphic way as the result of numerous impacts of photons bombarding the illuminated surface.

The magnitude of the force acting on a body is determined by the variation of its momentum per unit time. If radiation falls on a body (the disk in Lebedev's experiment),

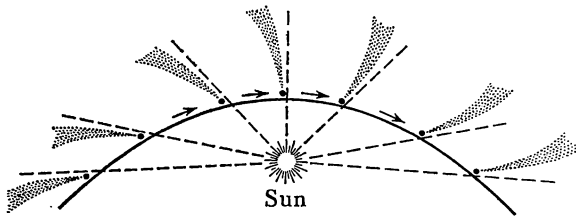


Fig. 38.2 Position of comet's tail with respect to Sun.

it transfers during time  $\Delta t$  a momentum  $\Delta p$ . This means that radiation acts on the disk with a force

$$E_p = \Delta p / \Delta t \quad (38.1)$$

If the surface completely absorbs all incident radiation (that is, if it is a black body), the variation of the momentum of radiation  $\Delta p$  will be equal to the momentum itself, which in the case of light is  $E/c$  ( $E$  is the energy of light; see Section 39-9). Since pressure

$$p = F_p / A \quad (38.2)$$

where  $A$  is the area of the illuminated surface, we get

$$p = \frac{E}{cA\Delta t} \quad (38.3)$$

For a body which completely absorbs all incident radiation,  $E$  is the energy falling on the body in the time  $\Delta t$ , and  $E/A \Delta t$  is the radiation energy falling on the body per unit area per unit time. The term for the latter is *intensity* of the wave. Denoting the intensity by  $J$ , we obtain the Maxwell formula for calculating the pressure exerted by light on the surface of a body which completely absorbs all incident radiation:

$$p = J/c \quad (38.4)$$

Lebedev's experiments proved this relation to be true.

### 38-3 The Thermal Effect of Radiation

The absorption of radiation is always accompanied by the transformation—partial or complete—of radiation energy into the internal energy of the body. Energy of all electro-

magnetic waves, from radiowaves to  $\gamma$ -rays, can be transformed into heat.

The thermal effect of radiation can be easily established through experiment by focusing solar rays with a lens on a piece of paper or a wooden surface or a piece of celluloid film. A charged spot soon appears on the paper or wood, while the film ignites. The energy concentrated in the beam of a powerful laser (see Section 38-18) is so high that even the best refractory metals evaporate in it. This method is used to make tiny holes in even the hardest materials, such as diamond.

The part played by solar radiation in phenomena occurring on the Earth is very great. The energy transported by radiation to the Earth is immeasurably greater than the total energy consumed by all industry. Solar rays bring 1370 J of energy per square metre of the cross section of the Earth just outside the atmosphere every second. This quantity is termed the *solar constant*

$$j = 1370 \text{ J/ (m}^2 \cdot \text{s)} = 1370 \text{ W/m}^2$$

#### 38-4 The Chemical Effect of Radiation

The chemical processes induced by radiation, *photochemical processes*, are of very great importance in nature, science and industry.

The term for one of the most important photochemical processes in nature, the assimilation by plants of atmospheric carbon dioxide, is *photosynthesis*. The leaves of plants, through the medium of chlorophyll excited by light, absorb carbon dioxide and liberate oxygen. This reaction is the cornerstone of the circulation of carbon and oxygen in nature: animals breathe out carbon dioxide and inhale oxygen; a reciprocal process activated by light takes place in plants.

Many chemical substances are decomposed by light. Such a reaction is of paramount importance for the production of visual sensations in men and animals. The retina contains about 120 million light-sensitive cells termed *rods*. The substance which fills these cells (termed *visual purple*) is decomposed by light, the decomposed products stimulating the ends of the nerves and thereby creating a visual sensation.

Besides the rods the retina contains about six million *cones*. These are light-sensitive cells of three types, each sensitive to one of the three primary colours—red, green and blue. The superposition of visual sensations from the nerve ends in those cells makes colour vision possible.

### 38-5 Photography

The formation of a photographic image involves the decomposition of molecules of silver bromide  $\text{AgBr}$  contained in the light-sensitive coating of a photographic plate by the incident radiation, the decomposition being accompanied by the liberation of particles of pure silver. The number of silver particles liberated depends on the time and intensity of irradiation of the photoplate. The greater the amount of radiation falling on a specific part of the plate, the greater the number of silver bromide crystals in which individual  $\text{AgBr}$  molecules are reduced to pure silver. These parts of the plate contain contrast parts of the virtual image invisible to the human eye.

The developer reduces to pure silver every silver bromide crystal in which one or more  $\text{AgBr}$  molecules have already been reduced to pure silver, the crystals containing only  $\text{AgBr}$  molecules being unaffected. Since silver is nontransparent, this means that when developed those parts of the plate which absorbed more radiation become darker. This method is used to obtain a negative image.

To obtain a photograph light-sensitive paper is placed under the negative and the latter is illuminated. Next the image on the paper is developed and fixed. This method is used to obtain a positive image.

The chemical effects of radiation can be well explained in terms of the quantum theory of radiation. The molecules increase their energy (are activated) by absorbing photons (quanta) and this sets off chemical processes in the substance. Quanta of small energy cannot activate the molecules and initiate chemical processes; their absorption results only in the substance being heated. The greater the energy of radiation quanta the greater their chemical activity. Accordingly, the chemical effects of radiation are the more pronounced the shorter its wavelength. For instance, ultraviolet rays exercise a great chemical effect on a photoplate, while red rays have no effect on a standard plate. Therefore such plates can be developed in red light.

### 38-6 External Photoelectric Effect

In 1887 Hertz discovered that the irradiation of a spark gap with ultraviolet light at high voltage applied across it facilitates spark discharge in air, that is, discharge takes place at greater gaps than in the absence of radiation. The



effects of radiation on electrical phenomena are termed *photoelectric effects*, or more briefly *photoeffect*.

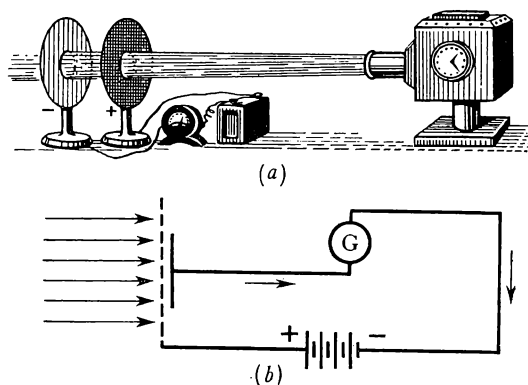
Photoeffect can be observed in the following experiment. If an electroscope containing a zinc plate instead of the ball is negatively charged ultraviolet irradiation will cause its rapid discharge. If, on the other hand, the electroscope carries a positive charge, irradiation will not affect its charge. Irradiation of the plate of a noncharged electroscope creates a small positive charge.

Similar experiments led to the conclusion that radiation leads to the escape of negative charges from metals. This effect is termed *external photoelectric effect*. It was later established that these charges ejected out of metals are electrons and that the external photoeffect is observed not only in irradiated metals but also in other solids, as well as in liquids and gases. Hence, the term applies to the ejection of electrons out of substances by incident radiation. (Explain why irradiation of the plate of a positively charged electroscope in the experiment described above does not cause a change in its charge.)

In the study of laws governing the photoeffect a great contribution was made by the Russian scientist Alexander G. Stoletov (1839-1896). In 1888 he repeated Hertz' experiments and established that high voltage was not essential for the photoeffect, since it takes place even if the voltage across the electrodes is quite small.

Stoletov constructed a device to obtain electric current with the aid of the external photoeffect (photocurrent) and to study the dependence of photocurrent on the intensity and wavelength of radiation. The apparatus in Stoletov's experiment is illustrated in Fig. 38.3. The radiation from

**Fig. 38.3** Stoletov's experiment: (a) apparatus (b) schematic diagram of experiment.



an electric arc passes freely through a positive mesh electrode and, falling on a negatively charged zinc plate (the negative electrode), knocks electrons out of it; the electrons rush to the grid and thereby set up a current, which is measured by a sensitive galvanometer.

Stoletov established that the photoeffect occurs most readily with violet rays and that the photocurrent is proportional to the intensity of the flux of those rays.

### 38-7 The Laws of External Photoelectric Effect

In studying the external photoelectric effect one should use, to obtain accurate results, electrodes of chemically pure materials in high vacuum to eliminate the effects of

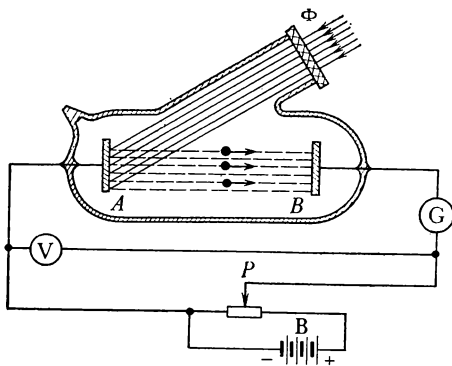


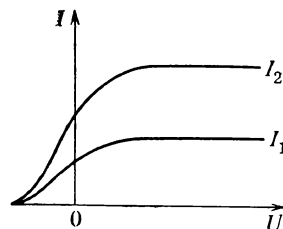
Fig. 38.4 Schematic representation of apparatus to study laws of external photoelectric effect (circles denote electrons).

air on the photocurrent and monochromatic radiation. The schematic diagram of such an arrangement is depicted in Fig. 38.4.

The voltage across the electrodes is measured with the aid of the voltmeter  $V$  and adjusted with the aid of the potentiometer  $P$ . Monochromatic light is made to fall on the negatively charged electrode  $A$  and the photocurrent is measured by the galvanometer  $G$ .

With a constant luminous flux  $\Phi$  an increase in the voltage at first causes the photocurrent to increase, but then it remains constant (i.e. no longer changes with the voltage (Fig. 38.5)). The maximum current obtainable with a constant luminous flux is termed *saturation photocurrent*. Obviously the voltage required to obtain saturation photocurrent should be such that all the electrons knocked out of the electrode  $A$  by the luminous flux reach the electrode  $B$ .

Fig. 38.5 Dependence of photocurrent on voltage.



Consequently, saturation photocurrent can serve as a quantitative measure of the photoeffect. Gradually increasing the luminous flux falling on the electrode  $A$  and measuring the photocurrent one can establish the *first law of external photoelectric effect*: the saturation photocurrent is directly proportional to the luminous flux falling on the electrode.

Decreasing the voltage with the luminous flux remaining constant, we observe that the photocurrent begins falling at low voltages, but even at zero voltage the current in the circuit does not vanish altogether. This means that the radiation falling on the electrode and knocking electrons out of it imparts kinetic energy to them.

The following method can be used to find this energy. Reverse the polarity of the battery  $B$ . This causes the electric field between the electrodes  $A$  and  $B$  to decelerate the electrons moving from  $A$  to  $B$ . Increasing the retarding field one can eventually stop the photocurrent altogether (see Fig. 38.5). In this case even the electrons which left the electrode at maximum speed are no longer able to overcome the retarding action of the electric field and do not reach electrode  $B$ . Denoting the minimum voltage at which the photocurrent is interrupted (the *cut-off voltage*) by  $U_{co}$ , the maximum speed of the photoelectrons by  $v_{max}$ , and the charge and the mass of the electron by  $e$  and  $m$  respectively, we can write

$$\frac{1}{2} m v_{max}^2 = e U_{co} \quad (38.5)$$

since in this case the maximum kinetic energy of the electrons will be equal to the work performed by them against the forces of the electric field along the path from  $A$  to  $B$ . By measuring the cut-off voltage  $U_{co}$  we can determine the maximum kinetic energy of the photoelectrons.

Such measurements led to the discovery of the second law of external photoelectric effect: the maximum kinetic energy of photoelectrons is independent of the intensity of radiation, being dependent only on its frequency  $f$  (or wavelength  $\lambda$ ) and on the electrode's material.

Illuminating the electrode with different monochromatic radiation we observe the maximum kinetic energy of the photoelectrons to decrease with the increase in wavelength, and the photoeffect to cease altogether at some wavelength. The maximum wavelength at which the photoeffect still can be observed is termed the *photoeffect threshold* for the specific material.

Experiments carried out with different materials led to the establishment of the *third law of external photoelectric*

*effect*: the photoeffect threshold is determined solely by the electrode material and is independent of the intensity of radiation.

### 38-8 Einstein's Photoelectric Equation

While the first law of *external photoelectric effect* can be explained on the basis of the wave theory of radiation, the second and third are in apparent contradiction with this theory.

Indeed, according to the wave theory an increase in the intensity of incident radiation, no matter what its wavelength, should result both in an increase in the energy of the photoelectrons and their number, that is, in the photocurrent, but in fact only the photocurrent rises. Next, it follows from the wave theory that the energy required to pull electrons out of a metal can be obtained from radiation of any wavelength provided its intensity is high enough. However, when, for example, a zinc plate is irradiated with yellow light of any intensity, no photoelectric effect is observed; at the same time ultraviolet radiation of negligible intensity produces the effect. All attempts to explain these properties on the basis of wave theory failed. But in 1905 Albert Einstein (1879-1955) demonstrated that the laws of photoelectric effect can be explained on the basis of the quantum theory.

We recall that an electron can leave the surface of any body, for example, of a metal, only if its kinetic energy is equal to, or exceeds, the work function (see Section 20-1). Let the radiation falling on the metal consist of photons with an energy  $hf$ . The electrons near the surface of the metal absorb the photons penetrating into the metal, acquiring their energy. The interaction of radiation with the substance in this case consists of an enormous number of elementary acts in each of which one electron completely absorbs one quantum (photon). If the magnitude of the energy of the quantum exceeds the work function, the electron is able to leave the metal. The major part of the quantum's energy is spent on the work function, the remainder comprising the electron's kinetic energy.

Obviously, the greatest kinetic energy will be that of electrons which absorb the quanta near the metal's surface and leave the metal without having lost energy in collisions with other particles. The mathematical expression for this

statement is *Einstein's photoelectric equation*:

$$hf = W + \frac{1}{2} mv_{max}^2 \quad (38.6)$$

or

$$\frac{hc}{\lambda} = W + \frac{1}{2} mv_{max}^2 \quad (38.6a)$$

The quantum theory explains the laws of photoelectric effect as follows. The increase in the intensity of monochromatic radiation causes an increase in the number of quanta absorbed by the metal and, consequently, in the number of photoelectrons; therefore, the photocurrent is directly proportional to the intensity of radiation (the first law).

It follows from (38.6) that the kinetic energy of photoelectrons depends only on the metal ( $W$ ) and on the frequency  $f$  (or wavelength  $\lambda$ ), that is, on the magnitude of the energy of the quanta, being independent of the radiation intensity (the second law).

If the magnitude of the energy of the quanta is less than the work function, the electrons will not leave the metal, no matter what the intensity of the radiation (the third law). The wavelength corresponding to the photoeffect threshold can be found from formula (38.6a) if the kinetic energy of the electrons is equated to zero:

$$\frac{hc}{\lambda_{th}} = W, \quad \text{or} \quad \lambda_{th} = \frac{hc}{W} \quad (38.7)$$

The values of threshold wavelength calculated with the aid of formula (38.7) agree well with experimental values. Experiments also confirmed that the kinetic energy of photoelectrons rises with the radiation frequency in full agreement with Einstein's equation (38.6). Experiments carried out not only with light but with X rays and  $\gamma$ -radiation as well, provided brilliant proof of the quantum theory of radiation.

### 38-9 Photocells Utilizing the External Photoelectric Effect

The external photoelectric effect can be used to convert radiation energy into electric. The device used for such a conversion is termed *photocell*. A description of the type of photocell utilizing the photoelectric effect, termed *photovoltaic cells*, follows.

The internal surface of an evacuated glass bulb is coated with a light-sensitive layer L with a small window W left

to let light into the bulb (Fig. 38.6). A metal ring R is mounted in the middle of the bulb. Electrodes are connected to the light-sensitive coating L and the ring R to connect the cell into a circuit. A drawing of such a cell is shown in Fig. 38.7.

Alkali metals are often used for the light-sensitive coating because of their low work function and consequent high sensitivity to visible light. Cells sensitive only to ultraviolet light are also manufactured.

Only a negligible fraction of the energy of radiation is converted into electric energy in photovoltaic cells, and for this reason they are not used as sources of electric energy. But such cells, actuated by visible or ultraviolet light signals, are used for the automatic control of electric circuits.

The advantages of such photocells are rapid response and the proportionality of the photocurrent to the radiation intensity, the latter quality making them useful for photometry. Their drawbacks are a weak current in the cell's circuit (which can, however, be amplified with the aid of vacuum tubes), insufficient sensitivity to long-wave radiation, fragility and comparatively high cost of manufacture.

To amplify the current in a photocell it is sometimes filled with rarefied gas ionized by the photoelectrons. Such photocells operate with larger voltages across the electrodes. The current in such cells is not proportional to illuminance.

Practical uses of photocells will be discussed in Section 38-13.

### 38-10 Internal Photoelectric Effect

In 1873 the British electrical engineers May and Smith, who tested an under-water cable, used selenium as an insulating material. In the course of tests May discovered that the resistance of selenium dropped upon illumination.

We recall that selenium is a semiconductor. There are some types of intrinsic semiconductors, which have very few free charge carriers (electrons and holes) in normal conditions and whose specific resistance is very great. However, the valence electrons in semiconductors are rather weakly bonded to the atoms and gaining excess energy are able to tear away from the atoms and become free. In an irradiated semiconductor the bound electrons absorb photons penetrating into it and become free.

Fig. 38.6 Schematic diagram of photocell using the external photoelectric effect.

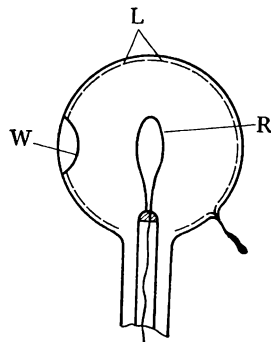
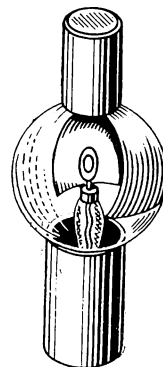


Fig. 38.7 Photocell.



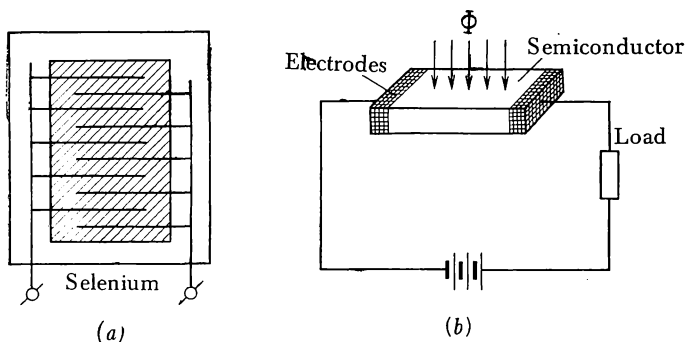
This results in an increase in the concentration of free charge carriers in the irradiated semiconductor and, consequently, in an increase in its conductivity. The term for the generation of free charge carriers in a semiconductor caused by irradiation is *internal photoelectric effect* (or simply *photoconductive effect*).

Take note of the principal difference between the photoemissive external photoelectric effect and the photoconductive effect: the former involves the ejection of electrons out of a substance; in the latter the electrons remain inside it. Since less energy is required to generate free charge carriers in a semiconductor than to knock them out of a substance, the internal photoelectric effect can be caused by radiation of greater wavelength than the photoemissive. Some semiconductors exhibit the effect in infrared light, which is of great practical importance. The term for the additional conductivity of semiconductors induced by radiation is *photoconductivity*. This effect is utilized in photoresistors and photocells.

### 38-11 Photoresistors

Devices whose resistance is changed by incident radiation are termed *photoresistors*. They are used for the control of electric circuits by means of light signals. In contrast to

**Fig. 38.8** Schematic representation of (a) photoresistor and (b) its connection into a circuit.



photocells photoresistors can be used in ac circuits, because their electrical resistance is independent of the direction of the current.

A photoresistor is made of a piece of semiconducting material of high sensitivity to light with a sufficient illuminated area (Fig. 38.8a). Since at room temperature the semiconductor's conductivity is very low, in the absence of

illumination a small (dark) current flows in the circuit (Fig. 38.8*b*). Illumination of the semiconductor decreases its resistance and the current in the circuit rises in proportion to the illumination.

Since the radiation penetrates the semiconductor only to a small depth and the changes in conductivity take place only inside a thin surface layer, there is little sense in making massive photoresistors. To manufacture photoresistors a thin layer of the semiconductor is deposited on an insulating substrate with electrodes previously deposited on it (Fig. 38.8*a*). The device is then protected with a transparent film.

The materials used in photoresistors include silicon (Si), selenium (Se), thallium sulphide ( $\text{Te}_2\text{S}$ ), bismuth sulphide ( $\text{Bi}_2\text{S}_3$ ), cadmium sulphide (CdS), etc. Each of these materials has its own peculiar properties determining its field of application. For example, different semiconductors have maximum photosensitivity in different wavebands. CdS has excellent photoelectric properties. It is sensitive only to radiation with a wavelength of  $0.5\text{ }\mu\text{m}$  and its specific resistance can decrease as much as a million times upon illumination.

The advantages of photoresistors include high photosensitivity, long life, small dimensions, low cost of production and facilities for choosing a photoresistor for the required waveband, specifically, for the infrared part of the spectrum.

Their drawbacks are the absence of direct proportionality between the current in the circuit and the illuminance, temperature dependence of resistance and slow response. The latter deficiency is due to the fact that the photoconductivity does not instantaneously drop to zero after the illumination has ceased, but decays with time. This is because free charge carriers generated by light vanish not instantaneously but after a time lag. This time lag may, by special treatment of the semiconductor, be brought down to  $10^{-8}\text{ s}$  (for silicon). Because of its persistence the photoconductivity is unable to keep up with rapid changes in illumination which take less time than its decay time.

### 38-12 Photocells Utilizing the Internal Photoelectric Effect

The photoconductive effect is utilized to convert radiation energy into electric energy in  $p$ - $n$  junction photocells. A device that has become quite popular for the conversion

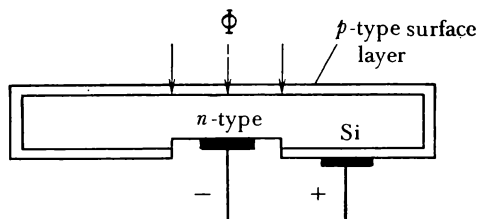


of solar energy into electric energy is the silicon photocell, known as the *solar battery*.

Solar batteries convert approximately ten per cent of the incident solar radiation into electric energy, this proportion exceeding the proportion of solar energy utilized in photosynthesis in plants.

One part of the solar battery (Fig. 38.9) is a chip of *n*-type silicon with a surface layer of *p*-type silicon of about one micrometre thick provided with contact electrodes for

Fig. 38.9 Schematic representation of solar battery utilizing the photoconductive effect.



external connection. We recall that all uncompensated charges are concentrated in the *p-n* junction region, the *p*- and *n*-regions themselves remaining neutral (see Section 24-4).

Illumination of the element's surface generates electron-hole pairs in the thin *p*-type layer, the majority of which, unable to recombine in this layer, reach the *p-n* junction and are separated there by the field: the electrons are drawn by the field into the *n*-region and the holes are rejected to the *p*-region. This means that an emf is generated across the electrodes whose magnitude can be as high as 0.5 V. When the electrodes are short-circuited the cell can produce a current of up to 25 mA ( $25 \times 10^{-3}$  A) per square centimetre of illuminated surface.

The maximum sensitivity of the silicon photocells lies in the green range, that is, in the waveband of maximum spectral density of solar radiation. This is one of the reasons for their rather high efficiency. Solar batteries mounted on artificial Earth satellites and on spacecraft supply the electric energy.

Germanium photocells are more sensitive to infrared radiation than those of silicon; therefore, they are more frequently used for operation in conjunction with artificial light sources. Other semiconductors are also used in photocells, for instance selenium, a thin layer of which is deposited on a metal. In this case a barrier layer is established

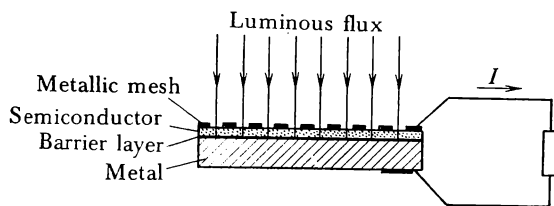


Fig. 38.10 Barrier photo-cell.

between the semiconductor and the metal, whose action is similar to that of a  $p$ - $n$  junction. Such photocells have become known as *barrier photocells* (Fig. 38.10).

### 38-13 Photocells in Science and Technology

The most widespread application of photocells is in movies for the reproduction of sound recorded on the sound track. Simultaneously with the filming of the frames sound

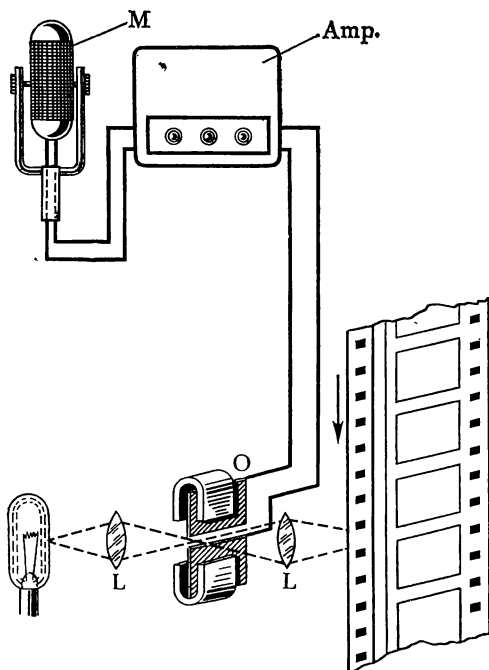
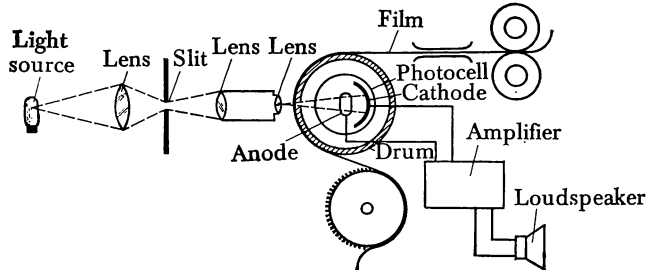


Fig. 38.11 Optical sound recording.

is recorded on the film in the form of semitransparent stripes of varying thickness or of dark spots covering various parts of the track.

A schematic diagram of an optical sound recording device is depicted in Fig. 38.11. The microphone *M* transforms sound vibrations into electric oscillations, which after being amplified in the amplifier *Amp.* passes through the blades of an optical "chopper" *O* placed close to each other between magnetic poles. The varying Ampere force sets the blades in

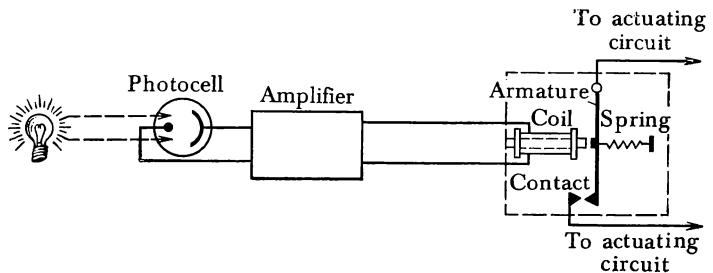
Fig. 38.12 Sound reproduction in a cinema.



motion, changing the gap between them and causing a corresponding variation in the amount of light falling on the sound track of the film on which the sound is recorded.

In the process of sound reproduction (Fig. 38.12) a narrow beam of light falls on the photocell through the sound track. The dark spots on the track absorb some of the luminous

Fig. 38.13 Photorelay.



flux. As the film moves, the magnitude of the luminous flux transmitted by the sound track changes continuously, causing changes in the current in the circuit of the photocell (or photoresistor). The current oscillations, amplified by the amplifier, are transmitted to the loudspeaker, which converts them into sound vibrations.

A device frequently used for the automatic control of various technological processes is the *photorelay* (Fig. 38.13). The photorelay consists of a photocell, a photocurrent amplifier and an electric relay. When the photocell receives ra-

diation, a current appears in the relay coil. The core in the coil is magnetized and attracts the armature against the force of the spring, closing the contact of the actuating circuit carrying high-power current that drives various mechanisms, instruments, etc.

In photorelays photoresistors are often used in place of photocells. When radiation falls on a photoresistor, its resistance drops and the current in the circuit increases.

There are numerous uses for photorelays. They switch beacons and street lights on and off. When the illuminance

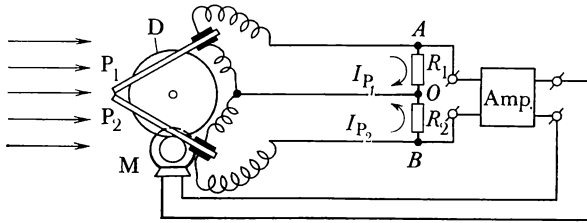


Fig. 38.14 Circuit diagram of astrocompass.

of the photocell drops below a specified level, the photorelay automatically switches on lighting, and when sunlight provides sufficient illuminance it switches it off. Photorelays stop machines in a paper plant or printing shop when the paper is torn, protects the worker from accidents, counts parts on an assembly line, controls the dimensions of parts, etc.

In science photocells are used to measure light intensity, luminance, illuminance (in photometers and in luxmeters), and to detect invisible radiation. In the military field photocells are used in self-aiming projectiles, for signal communications and for location with invisible rays. In communications photocells are used in the phototelegraph to transmit static images, in phototelephones operating without wires on infrared rays, etc.

Consider one more example of the use of photocells—in the astrocompass, an automatic device for orientation by the Sun and the stars. The circuit diagram of an astrocompass is depicted in Fig. 38.14. The disk  $D$  carries two identical selenium photocells  $P_1$  and  $P_2$  mounted at an angle to each other. They are connected to identical resistors  $R_1$  and  $R_2$ . The voltage from these resistors,  $U_{AB}$ , is applied to the input of the amplifier  $\text{Amp.}$  and after amplification is supplied to the motor  $M$ , whose direction of rotation depends on the polarity of the voltage across its terminals.

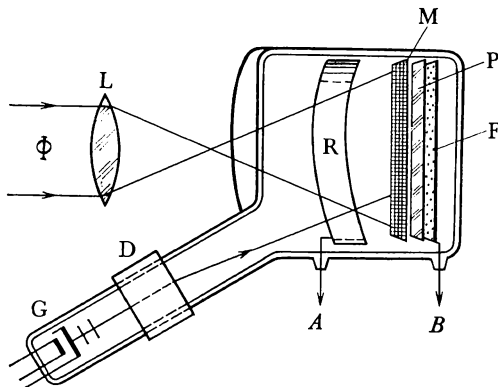
When the photocells make identical angles with the solar rays, that is, when their illumination is identical, then  $I_{P_1} = I_{P_2} = U_{AO} = U_{BO}$  and  $U_{AB} = 0$ . When the astrocompass is turned, for instance when the aeroplane changes course, one of the photocells receives more radiation, the photocurrents are no longer equal, and a voltage of the appropriate sign will appear at the input (and, consequently at the output) of the amplifier. The motor will start turning the disk carrying the photocells in the direction of the escaping Sun. When the illumination of the photocells is again equal, the voltage across the motor is zero and it stops. In this way the astrocompass "follows" the Sun. The motion of the Sun across the horizon due to the diurnal rotation of the Earth is compensated for by rotating the reference scale with the aid of a clockwork mechanism.

The high sensitivity of photocells makes orientation by the stars also possible. The astrocompass is turned to some star of sufficient brightness and it will continue to follow it, indicating its direction. Such devices are used in place of the magnetic compass in polar aviation and in spacecraft.

### 38-14 Television

*Television* is the term used for the long-distance wireless transmission of images. Television is achieved in the following way. Light signals are transformed into electrical

**Fig. 38.15** Schematic representation of iconoscope.

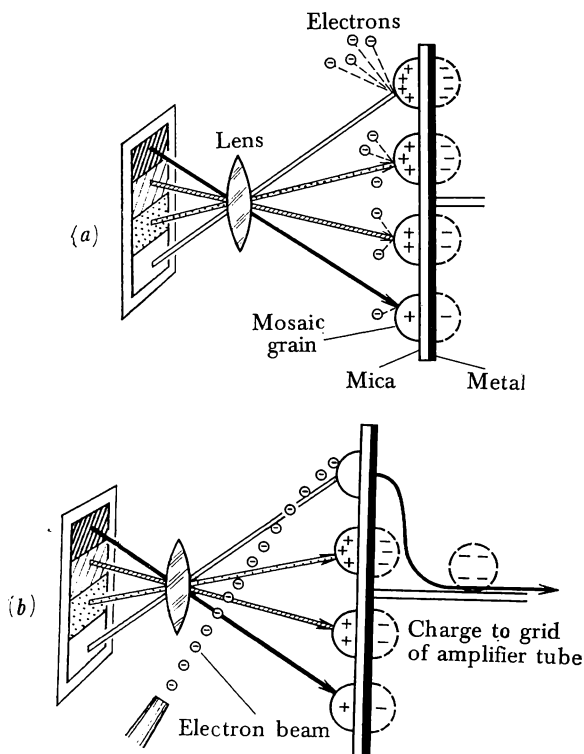


signals, then transmitted by electromagnetic waves across great distances, and finally received by the television aerial and again transformed into light.

The transformation of light signals into electric signals takes place in a cathode-ray tube called an *iconoscope* (Fig.

38.15). It consists of a mosaic capacitor M, a collector ring R, an electron gun G and a deflection system D which controls the electron beam. The iconoscope is evacuated to a high vacuum like all cathode-ray tubes. The lens L forms the image to be transmitted on the mosaic capacitor.

The design of the mosaic capacitor is described below. Grains of caesium silver—tiny photocells highly sensitive to visible radiation—are deposited on one side of a very thin mica plate P, the opposite side of which is coated with a so-



**Fig. 38.16** Schematic representation of operation of mosaic capacitor: (a) light knocks electrons out of mosaic grain; (b) electron beam discharges mosaic grain and negative charge on opposite side of plate goes to grid of amplifier tube.

lid conducting film F. If one of the photocells acquires a positive charge, a negative charge appears on the other side of the mica plate. Hence, the entire plate consists of a multitude of microscopic capacitors.

When an image of an object is formed on the mosaic, light rays of different intensity fall on the individual photocells (Fig. 38.16a), knocking electrons out of each of them in proportion to the intensity of the rays. This causes the photocells to acquire positive charges and the conductive sur-

face opposite them negative charges. This produces a virtual image on the mosaic capacitor in the form of a distribution of positive charges of varying magnitude on the photocells.

If an electron beam is directed onto the photocell, it neutralizes the positive charge on the photocell, liberating the negative charge on the opposite side of the plate and enabling it to flow via a wire to the grid of an amplifier tube (Fig. 38.16b).

Now imagine that the electron beam starts moving along the lower fringe of the mosaic surface and, coming to its edge, rises by a fraction of a millimetre and again starts moving in the horizontal direction, scanning line after line until it reaches the top fringe of the mosaic. When the beam touches a charged photocell, an electric pulse of a magnitude proportional to the charge on the photocell appears in the circuit of the mosaic electrode. Such pulses applied to the grid of an amplifier create current pulses of the same shape in the anode circuit of the tube. These are then fed to a UHF transmitter. The signals constitute the scanned image and are termed *videosignals*.

As the electron beam runs across the mosaic surface, it erases the virtual image, but light rays immediately recharge the photocells and the whole process of transmitting the signals is repeated as the electron beam scans the surface again. In the Soviet TV system the electron beam makes 625 lines on the mosaic surface, repeating the process 50 times per second. Hence, 50 frames are transmitted per second.

The metal ring R of the iconoscope (see Fig. 38.15) is designed to collect the electrons knocked out of the photocells. The wire A is connected from the ring to the cathode of the amplifier tube, and the wire B to the grid, forming the input circuit.

A TV station sends three types of signals: videosignals, signals controlling the motion of the electron beam (synchronization signals) and audio signals. The three signals are received by the aerial of the TV set, are amplified and then separated. The videosignals are sent to the TV set's cathode-ray tube, the *television tube*, where they control the intensity of the electron beam. The synchronization signals are sent to the scanning generator, synchronizing the motion of the electron beam in the tube with the motion of the beam in the iconoscope.

The image on the tube's screen is formed because the electrons in its electron beam excite the luminescence of the compound covering the entire surface of the screen. The luminous intensity of a spot on the screen depends on the

intensity of the electron beam at the moment it strikes the spot. In this way is the transmitted image reproduced. The audiosignals are sent to a radioreceiver of conventional design.

### 38-15 Bohr's Atom Model

As far back as 1885 the Swiss physicist J.J. Balmer (1825-1898) found a definite regularity in the position of the lines of the hydrogen spectrum and demonstrated that the wavelengths corresponding to lines in the visible part of the spectrum can be computed with the aid of a formula which is now written in the form

$$\frac{1}{\lambda} = R \left( \frac{1}{2^2} - \frac{1}{m^2} \right) \quad (38.8)$$

There is a spectral line to correspond to each integer  $m$  (above 2) in this formula, while  $R$  is a constant termed

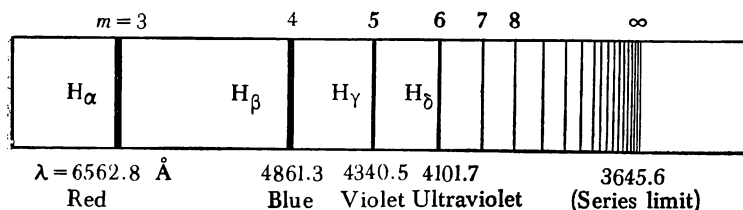


Fig. 38.17 Spectral lines of Balmer series (notation of individual lines is shown).

*Rydberg's constant:*  $R = 10\,967\,758\text{ m}^{-1}$ . All the lines of the hydrogen spectrum satisfying the relation (38.8) are termed *Balmer series* (Fig. 38.17).

Subsequently additional series of the hydrogen spectrum were discovered. These can be obtained if the 2 in (38.8) is substituted by an integer  $n$  so that the formula assumes the form:

$$\frac{1}{\lambda} = R \left( \frac{1}{n^2} - \frac{1}{m^2} \right) \quad (38.9)$$

Further experiments have demonstrated that formula (38.9) for integers  $m$  and  $n$  such that  $m$  exceeds  $n$  describes all the lines of the spectrum of atomic hydrogen obtainable in experiment. The lines corresponding to  $n = 1$  lie in the ultraviolet part of the spectrum and make up the *Lyman series*, while those corresponding to  $n = 3$  are in the infra-red part and make up the *Paschen series*.

Similar regularities in the position of spectral lines, although of a more intricate pattern than in the case of the



hydrogen atom, were also established for other atoms. Attempts to explain these experimental results on the basis of classical physics failed. Rutherford's nuclear model of the atom, far from being able to explain the spectral regularities, in terms of classical physics, could not even explain the existence of the spectral lines, that is, lines of definite wavelength emitted and absorbed by the atoms. It was not even possible to explain the stability of the atom itself.

Indeed, an electron spinning with a centripetal acceleration around the nucleus which attracts it should continuously radiate electromagnetic waves, losing energy in the process just like any accelerated charge. The radiation frequency of the electron would have to change continuously as the electron approached the nucleus with every revolution and eventually fall onto it. This is not what actually happens.

To explain such contradictions between experimental findings and classical physics scientists had to concede that the laws of classical physics were applicable to atoms only to a limited extent. The first to take this most important and bold step was the Danish physicist Niels Bohr (1885-1962), who combined quantum mechanical ideas with Rutherford's nuclear model of the atom. In 1913 Bohr developed his theory of the hydrogen atom, which agreed with all experimental data known at the time.

According to Bohr's ideas the atom cannot have an arbitrary energy. Every atom has *discrete* energy values—a series of definite energy values it can possess. An atom cannot exist in states with intermediate energies others than those of the allowed values. Those energy values allowed to the atom by nature are termed *energy levels*. It follows from the above that of all the electron orbits feasible from the point of view of classical physics only those are in fact possible which correspond to one of the atomic energy levels. Such a selection of allowed electron orbits became known as *quantization* of the orbits. Three postulates form the basis of Bohr's theory.

(1) Electrons can move about the atomic nucleus only in definite orbits corresponding to one of the atomic energy levels.

(2) When an electron moves in an allowed orbit, the atom is in a stationary state, that is, neither emits nor absorbs energy.

(3) When an electron jumps to another orbit closer to the nucleus, the atom emits a quantum of energy (a photon) in the form of radiation of a frequency determined by Planck's formula (31.1).

Thus, Bohr did not reject the laws of classical physics, but placed limitations on them in the form of quantization of orbits.

It follows from Bohr's postulates that the magnitude of the quantum's energy emitted by the atom in the process of transition from one stable state to another is equal to the difference in the atom's energy in those two states:

$$hf = E_m - E_n, \quad \text{or } f = E_m/h - E_n/h$$

This can be written in the form

$$\frac{1}{\lambda} = \frac{1}{hc} (E_m - E_n) \quad (38.10)$$

Bohr calculated the radii of the allowed orbits of the electron in the hydrogen atom and the corresponding energy levels. The allowed energy values of an atom turned out to be inversely proportional to the squares of the natural numbers (1, 4, 9, 16, etc.). As a result formula (38.9) was obtained from formula (38.10) and an expression for Rydberg's constant was obtained:

$$R = \frac{e^4 m}{8h^3 c \epsilon_0^2} \quad (38.11)$$

where  $m$  is the electron mass and  $e$  the electron charge. The theoretical value of  $R$ , calculated from formula (38.11) proved to be in excellent agreement with the experimental value.

Thus, Bohr's theory explained the entire spectrum of the hydrogen atom with remarkable accuracy. Although Bohr's theory as applied to more complex atoms of other elements did not produce quantitative results, it helped to get an insight into the nature of atomic spectra and to explain in general terms the laws which they obey.

Subsequently quantum mechanics was developed, with Bohr's theory included as a specific case. Note in addition that Bohr's postulates are a natural corollary of the equations of quantum mechanics.

Not only did the works of Bohr remove physics out of a deadlock, but they also opened up prospects for the rapid development of the new science of atomic physics.

### 38-16 The Quantized Atom

According to Bohr's theory an electron moving in the allowed orbit closest to the nucleus is in the *ground state* corresponding to its stable equilibrium position. The atom

can remain in the ground state for an indefinite time because this state corresponds to the lowest energy level available to the atom.

An electron moving in some other allowed orbit is in an *excited state*, which is less stable than the ground state. After some time (usually of the order of  $10^{-8}$  s) the atom spontaneously goes over from the excited to the ground state, emitting an energy quantum in the process.

Vice versa, the transition of an atom from the ground to the excited state involves an increase in its energy and can therefore take place only through an external influence on the atom, for example, when the atom absorbs a photon or collides with another atom or electron.

An atom in the ground state can absorb only a portion of the energy needed for its transition to one of the possible excited states. Therefore, in transition to a higher energy level the atom can absorb only one whole energy quantum.

The only exception is when the external action imparts to the atom an energy exceeding that required for its ionization. In this case a part of the external energy is spent on ionizing the atom, and the rest is transmitted to the electron, knocked out of the atom, in the form of kinetic energy. This can assume arbitrary values.

The energy of the quanta is usually expressed in electron volts. The term *electron volt* (eV) applies to work performed by an electric field in moving an electron between points with a potential difference of 1 V.

Since the expression for work is  $W = eU$ , it follows that

$$1 \text{ eV} = 1.6 \times 10^{-19} \text{ C} \times 1 \text{ V} = 1.6 \times 10^{-19} \text{ J}$$

To ionize an atom of hydrogen in the ground state an energy of 13.54 eV must be imparted to it.

Figure 38.18 depicts a diagram of a hydrogen atom showing five possible electron orbits. The radius of the orbit closest to the nucleus and corresponding to the ground (unexcited) state of the hydrogen atom was calculated by Bohr to be equal to  $0.53 \times 10^{-10}$  m.

In the transition of an electron of an atom in an excited state to an orbit closer to the nucleus the atom emits a quantum of energy in the form of radiation of a definite wavelength. The electron can go over, for instance, from the fifth orbit directly to the first or to any intermediate orbit.

Hence, it follows from Bohr's theory that  $m$  in formula (38.9) is the number of the orbit from which the electron transition takes place and  $n$  is the number of the orbit where the electron settles as the result of the transition.

Each of the transitions from one level to another is accompanied by the emission of energy quanta of different magnitude.

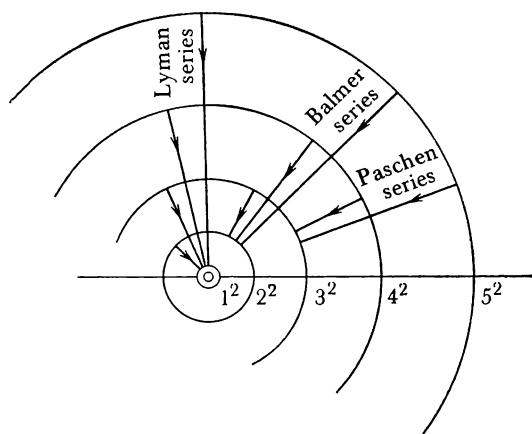


Fig. 38.18 Five allowed electron orbits in hydrogen atom (arrows indicate possible transitions).

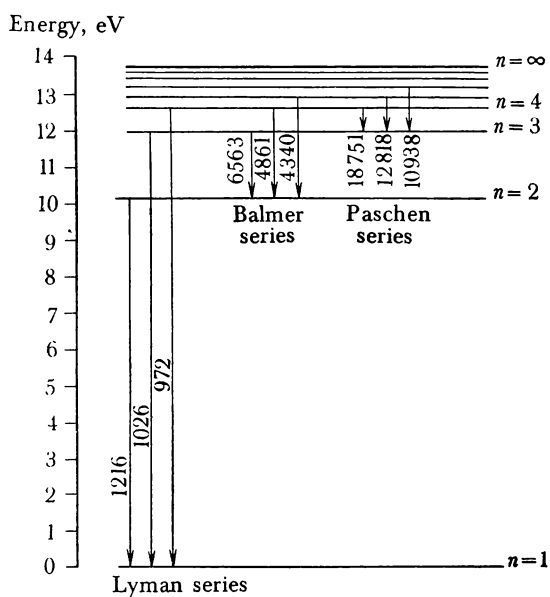


Fig. 38.19 Energy levels of hydrogen atom.

Figure 38.19 shows the energy levels of the hydrogen atom corresponding to different electron orbits. It can be seen that the transition of an electron from any of the higher orbits to the first results in the liberation of much greater

energy than the transition to the second orbit. This explains why the Lyman series lies in the ultraviolet part of the spectrum and the Balmer series in the visible part. To sum up:

(a) a free atom absorbs and emits energy only in whole quanta;

(b) in the course of transitions to excited states the atom absorbs only such quanta which it can itself emit.

The last statement means that free atoms absorb only such rays as they themselves can emit. This is why the positions of the lines of the absorption spectra coincide with those of the lines of the emission spectra (see Sections 37-9 and 37-10).

The exceptional constancy of atomic radiation frequencies was utilized for the definition of a new standard for the base unit of time, the second. By international agreement one of the frequencies of the caesium-133 atom was chosen for this purpose and at present the *second* is defined as a time interval in which a definite number (9192 631 770) of oscillations corresponding to this frequency takes place.

The energy-level diagram of the hydrogen atom considered above is the simplest of such diagrams. The more electrons an atom contains the more complex is the diagram of its energy levels and its spectrum. Thus, the spectrum of iron contains several thousand lines.

Molecules have still more complex spectra. The energy of a molecule is made up of three parts: the energy of the electrons, vibrational energy of the atomic nuclei, and the energy of rotation of the nuclei about their common centre of masses. All these parts are discrete and their variation obeys quantum laws.

Various combinations of these three parts produce an enormous number of molecular energy levels. Obviously, the number of possible transitions from one energy state to another is also very great. This leads to the formation of molecular band spectra, every band of which is made up of many closely spaced lines.

In solids and liquids, where the particles strongly interact with one another, the energy of each particle includes also the energy of its interaction with other particles. Since the interaction energy can assume quite different values, instead of separate energy levels continuous bands of possible energy states are formed. Because of this the magnitude of the radiation quanta can also be quite different and a continuous emission spectrum is obtained. The spectrum of thermal radiation belongs to the same type since its properties are determined by the temperature and depend little on the structure of the substance and its particles.

### 38-17 Luminescence

Any body heated to a high enough temperature begins to emit radiation. Such radiation is termed *thermal*, since its intensity and spectral composition depend mainly on the temperature of the luminous body.

However, frequently one is able to observe the luminosity of bodies at such low temperatures that there are no visible rays in their thermal radiation. Such luminosity is always due to the absorption by the body of energy which was not transformed into heat. If a measurable time interval lapses between the moment the body has absorbed energy and the moment it emits it in the form of radiation independent of thermal radiation, the term for such radiation is *luminescent radiation* and for the phenomenon itself *luminescence*.

Absorbing the appropriate energy the particles of a *luminophor* (a substance capable of luminescence—molecules, atoms, or ions) move to an excited state, in which state they remain for a definite time (depending on the substance, from  $10^{-9}$  s to several hours). Returning to the normal state they emit luminescent radiation. There are several types of luminescence, distinguished by the method of excitation.

The terms for the luminosity of a rarefied gas (see Section 23-4), excited by electric current passing through it, is *electroluminescence*. Electroluminescence is also observed in semiconductors and is utilized in light-emission diodes. When a forward current flows through a *p-n* junction, an intense recombination of electrons and holes takes place, accompanied by the emission of quanta of radiation. In this case a transformation of electrical energy into luminous energy takes place, a process reciprocal to the photoconductive effect. Silicon light-emission diodes are sources of infrared radiation, and light-emission diodes made of silicon carbide (SiC) and gallium phosphide (GaP) emit visible light.

Luminescence accompanying the absorption by a body of luminous radiation is termed *photoluminescence*. As a rule photoluminescence of solids and liquids entails the emission of radiation of longer wavelength than that of absorbed radiation. Normally ultraviolet radiation is used for excitation, the resulting photoluminescent radiation being in the visible part of the spectrum. Thus, a sort of transformation of radiation takes place. This property of photoluminescence was discovered in 1852 by the British scientist Sir George G. Stokes (1819-1903) and is now termed *Stokes' rule*. the photoluminescence spectrum is displaced in the direction of longer waves as compared to the spectrum of absorbed radiation.

The quantum theory explains this rule as follows: a molecule (or atom or ion), having absorbed a quantum of radiation  $hf_0$  and moved to an excited state, can lose a part of the energy it gained to other molecules in the process of thermal motion, emitting the remaining energy in the form of a quantum  $hf$ . Denoting the energy lost by the molecule by  $E$ , we obtain:

$$hf = hf_0 - E \quad (38.12)$$

Accordingly, the frequency of luminescent radiation is lower than that of absorbed radiation, the wavelength being correspondingly greater. A very important point is that the spectrum of luminescent radiation is practically independent of absorbed radiation, thus being a characteristic feature of the luminophor. This fact is utilized in luminescent analysis to determine the composition and degree of purity of substances and to detect nonuniformity of structure. The sensitivity of this analytical method is very high; usually the luminescence of a substance can be observed if it is present in concentrations of  $10^{-7}$  to  $10^{-9}$  g/cm<sup>3</sup>.

The term for the time after illumination during which the luminescence of a substance can still be observed is *luminescence decay time*. According to the length of the decay time luminescence is divided into fluorescence and phosphorescence. If the decay time of luminescence is so small that it ceases practically the moment the irradiation of the substance has ceased, the term for it is *fluorescence*. If the decay time of the luminescence is appreciable (sometimes more than 24 hours), the term for it is *phosphorescence*. Many liquids and gases exhibit fluorescence, phosphorescence being the privilege of solids.

Crystalline substances exhibiting intensive and durable phosphorescence are termed *phosphor crystals*. They include various salts containing tiny amounts of dopants of specific substances termed *activators*. For instance, zinc sulphide emits intense green phosphorescent light when activated by copper atoms. Many glasses containing some luminescent substance, such as uranium compounds and rare earth elements, exhibit phosphorescence.

Phosphor crystals are used to detect X rays and ultraviolet radiation. Absorbing such radiation screens coated with a phosphor-crystal layer emit visible light. Phosphorescent screens can also be used to detect infrared radiation. Actually, infrared radiation reduces the phosphorescence decay time of a luminophor, causing the luminous screen to fade rapidly.

Luminescence is widely used in cathode-ray tubes, whose screens are coated from the inside by a luminophor emitting light when bombarded by electrons. The term for such luminescence is *cathode luminescence*.

In conventional television tubes phosphor crystals containing a mixture of zinc and cadmium sulphide activated by silver and emitting a bluish light are used. The screen of a colour television tube has a layer of three different types of phosphor crystal grains, emitting red, green and blue light arranged in a regular order. They are excited by three separate electron beams. The intensity of the beams is controlled by vide signals transmitted from three transmitting tubes, each provided with a red, green or blue light filter.

Luminescence is widely used for light sources. Fluorescent lamps use the electroluminescence of rarefied gases or vapours. In "daylight" lamps the electroluminescence of mercury vapour produces ultraviolet radiation. (This is why mercury lamps with quartz bulbs are used as sources of ultraviolet light.)

The luminophor covering the inside surface of a "daylight" lamp's tube absorbs ultraviolet radiation and in the course of phosphorescence emits visible light. The chemical composition of the luminophor is chosen so as to make the spectral composition of the lamp's radiation similar to that of daylight. Such lamps are four to five times more economical than incandescent lamps.

Luminescent paints not only scatter light of a definite colour but also transform into such light the absorbed radiation of other colours, this being the reason for their apparent luminosity. They are used to create colour effects in theatres, in advertizing, to paint beacons, bright stripes on locomotives, etc.

### 38-18 Lasers and Masers

A new branch of physics, quantum electronics, was born in 1955, whose evolution resulted in the appearance of quantum oscillators—masers and lasers (short for *microwave and light amplification by stimulated emission of radiation*).

The quantum oscillator is a source of coherent electromagnetic radiation of a definite frequency and of high directivity. The maser operates in the microwave band and the laser in the visible and infrared parts of the spect-



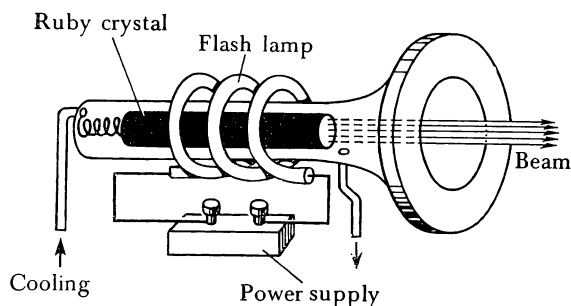
rum. The first quantum oscillators were made by the Soviet physicists Nikolai G. Basov and Alexander M. Prokhorov.

The radiation in the quantum oscillator is produced, as in the case of normal luminescence, by excited atoms or molecules which emit quanta as they return to their ground states. If such transitions occur spontaneously (as in the case of normal luminescence), the photons emitted fly in different directions and the corresponding waves have arbitrary phases. This means that the radiation in this case is incoherent and nondirectional.

However, the excited atom (or molecule) can give up its energy also in the act of *induced*, or *stimulated*, radiation, when it interacts with a similar photon radiated by another kindred atom. This interaction is of the resonance type and results in the emission of a similar new photon flying in the same direction, with its associated wave being in-phase with the wave of the origin photon. These photons stimulate the radiation of other excited atoms, and so on. Thus, instead of the usual attenuation of light caused by absorption its amplification takes place.

For powerful stimulated emission to be possible the atoms must remain in an excited state for some time, that

**Fig. 38.20** Schematic representation of ruby laser.



is, this state should be stationary. It is also necessary that most of the atoms be in the excited state. Such conditions are created in quantum oscillators.

By way of an example, let us consider the principle of the operation of an optical quantum oscillator, the ruby laser (Fig. 38.20) developed in 1960. It consists of a synthetic ruby crystal (aluminium oxide doped with chromium) in the shape of a rod whose ends are strictly parallel, finely polished and coated with silver, the left-hand end being nontransparent and the right-hand (the output) semitransparent.

The luminous radiation of the laser is produced by the chromium atoms excited with the radiation of a flash lamp—a powerful impulse gas-discharge tube made in the shape of a spiral wound around the crystal. The powerful flash excites most of the chromium atoms.

Suppose some excited chromium atom spontaneously emits a photon in the direction of the rod's axis. This photon stimulates the emission of other chromium atoms, the photons emitted by them stimulate the emission of new atoms, etc., and in this way an avalanche of photons develops. Since the waves corresponding to these photons are in-phase, an electromagnetic wave with a continuously increasing amplitude is generated.

Arriving at the end mirror it is reflected and turns back. As a result of multiple reflections a standing wave with a rapidly rising amplitude is formed, an integral number of half-waves being contained in the ruby crystal which thus acts as a resonator. Some of the radiation falling on the semitransparent mirror at the output end of the crystal passes outside, forming an exceptionally powerful monochromatic coherent radiation termed the *laser beam*.

Radiation due to photons moving at some angle to the crystal's axis cannot experience multiple reflections from its ends and, consequently, its total amplification is low. This explains the high directivity of laser radiation.

In the period of about a millisecond all excited chromium atoms return to the ground state and laser radiation ceases. The ruby laser emits short but very powerful flashes of red light.

A laser in operation liberates a great amount of heat, and cooling is required for it. Other solids besides the ruby are being used as active laser materials, such as some glasses doped with rare earth elements, as well as gases such as argon, nitrogen, a mixture of helium and neon.

In *gaseous lasers* radiation is emitted by a rarefied gas whose atoms are excited by a high-frequency electric current. The gaseous lasers operate continuously. Their radiation is not so powerful as that of solid-state lasers, but, on the other hand, it is even more directional and monochromatic.

A drawback of lasers is their low efficiency (less than one per cent). However, semiconductor lasers have now been developed which are, in effect, light-emission diodes operating at enormous current densities.

Lasers are being used in many fields of science and technology. A directional laser beam is used in building tunnels, laying pipelines, in construction, navigation and in defense

(for aiming guided missiles). The most refractory metals evaporate in a focused laser beam. This phenomenon is utilized for making very fine holes in ceramics, superhard alloys, diamonds and semiconductor materials. The perfect directionality of laser radiation is used for locating the Moon, Venus and Mars.

The coherent laser ray, like any other electromagnetic wave, can be used to transmit information. Since the volume of information a wave can carry increases with its frequency, the laser wave is able to transmit many thousand times more information than a radiowave. Photocells and photoresistors act as receivers of laser radiation. Because of the high absorption of the laser beam in the atmosphere, ground laser communication systems have to make use of waveguides (optical fibres), but in outer space the great advantages of laser systems can be exploited without limitations.

## 39 The Fundamentals of Special Relativity Theory

### 39-1 Relativity in Classical Mechanics

When describing physical phenomena we always make use of some reference frame. For example, we usually consider the motion of bodies with respect to the Earth, that is, we assume the Earth to be at rest. Sometimes we accept as the reference frame the floor and the walls of the laboratory which may be at rest with respect to the Earth, but which also may be in motion, for instance, if it is on a train or a ship or in a spacecraft.

Normally the Cartesian coordinate system and a clock are used to locate a point in space (Fig. 39.1). The position of a point at a specified moment of time  $t$  in such a system is determined by the three coordinates  $x$ ,  $y$ ,  $z$ , this fact being recorded for points  $A$  and  $B$  in the following form:  $A(x_1, y_1, z_1)$  and  $B(x_2, y_2, z_2)$ .

Galileo demonstrated that on the Earth the law of inertia is valid in practice. A reference frame in which the law of inertia is valid is called *inertial*.

It was established that in a laboratory moving in a straight line at a constant speed with respect to the Earth all mechanical phenomena take the same course as in a laboratory that is at rest with respect to the Earth and that, for

this reason, the law of inertia also holds in such a laboratory. This means that there are a lot of inertial reference frames, since if there is at least one such frame, any other frame moving in a straight line at a constant speed with respect to it will also be inertial.

Galileo introduced into classical mechanics the *principle of relativity*, to the effect that there are no mechanical experiments that could help us find out whether an inertial

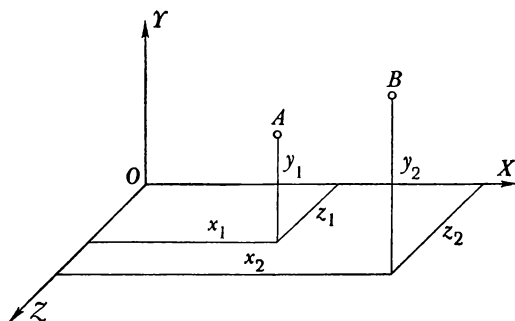


Fig. 39.1 Cartesian coordinate system: the position of particle in space is determined by three coordinates.

frame is at rest or whether it moves in a straight line at a constant speed. In other words, the laws of mechanics have the same form in all inertial frames and therefore none have any advantage over the others; any one of them can be assumed to be at rest and used to describe mechanical phenomena.

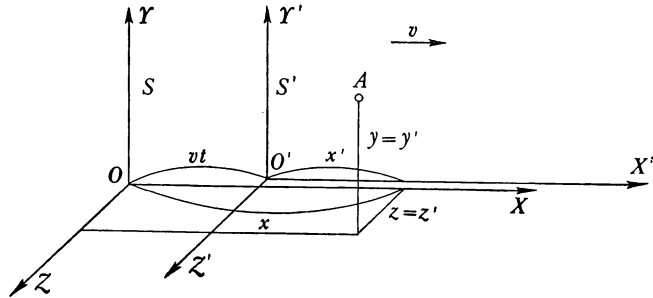
It should be noted that the reference frame<sup>\*</sup> of the Earth can be regarded as inertial only as an approximation, since the law of inertia is actually not strictly obeyed on the Earth because of its rotation about its axis. For instance, bodies falling to the ground are deflected eastwards, a pendulum changes its plane of oscillation (Foucault's experiment), etc.

The reference frame associated with the Sun is much closer to a true inertial reference frame. Strictly speaking, this frame is also noninertial since the Sun takes part in motion about the centre of the Galaxy. Hence, the question whether a chosen reference frame is inertial or not demands an experimental answer. If no measurable deviations from Newton's laws for an inertial frame are observed in experiment, the chosen reference frame can be regarded as an approximately inertial one.

### 39-2 Galilean Transformations

Imagine ourselves observing and describing some mechanical phenomenon, for instance the motion of a particle in space, from an inertial reference frame  $S$  at rest with respect to us. We want to know how an observer in another reference frame,  $S'$ , moving in a straight line at a constant speed  $v$  with respect to the first, would describe this phenomenon. Would there be any relationship between the formulae describing the motion of the particle in both reference frames? In other words, can one, knowing the formulae describing the mechanical motion of a particle with respect to some reference frame with the aid of some simple mathematical transformations obtain formulae describing this motion with respect to another frame? The answer is yes and the term for the relations to be used is *Galilean transformations*. Let us find these transformations.

Fig. 39.2 Coordinates of particle in stationary and moving reference frames.



If an inertial frame  $S'$  moves in a straight line with respect to the frame  $S$  at a constant speed  $v$ , a choice of orthogonal coordinate systems in both frames is possible.  $XYZ$  in the frame  $S$  and  $X'Y'Z'$  in the frame  $S'$ , such that the axes  $X$  and  $X'$  coincide with one another and with the direction of motions and that the axis  $Y$  is parallel to the axis  $Y'$  and the axis  $Z$  is parallel to the axis  $Z'$  (Fig. 39.2).

Let the primed system move in the positive direction  $X$  (i.e. to the right in Fig. 39.2) at a constant speed  $v$  and let the origins  $O$  and  $O'$  coincide at the moment  $t_0 = 0$ . Let a point  $A$  be specified in space at rest with respect to  $S$ . In such a case the point moves with respect to the second frame  $S'$  and its coordinates are variable.

If the coordinates of this point in the first system are  $(x, y, z)$  and in the second system  $(x', y', z')$ , obviously  $y' = y$  and  $z' = z$ . The coordinate  $x'$ , on the other hand, depends on time, since the primed system moves at a speed

$v$ . At the moment  $t_0 = 0$  the coordinates of the point  $A$  in both systems coincide, since at that moment their origins  $O$  and  $O'$  coincide.

During the time  $t$  the moving system covers a distance equal to  $vt$ . Hence,  $x' = x - vt$  and we can write

$$x' = x - vt, \quad y' = y, \quad z' = z, \quad t' = t \quad (39.1)$$

Formulae (39.1) are Galilean *direct* transformations. The last equation shows that the course of time in both frames is identical, this, until the appearance of Einstein's works, being regarded as an obvious fact.

If the point  $A$  remains at rest in the primed system, its coordinates in the unprimed system moving with respect to the primed at a speed of  $-v$  (to the left in Fig. 39.2) will be expressed as follows:

$$x = x' + vt, \quad y = y', \quad z = z', \quad t = t' \quad (39.2)$$

These are Galilean *inverse* transformations.

Suppose now the particle moves with respect to both coordinate systems. For the sake of simplicity we assume

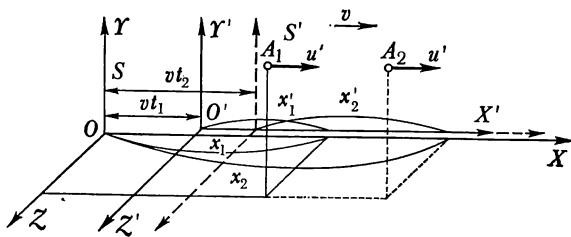


Fig. 39.3 Coordinates of moving particle at various moments of time.

the particle to move parallel to the  $O'X'$  axis in the positive direction of  $X'$  at a speed  $u'$  with respect to the primed system. What will its speed  $u$  be with respect to the unprimed system?

Let the particle occupy position  $A_1$  at time  $t_1$  and position  $A_2$  at time  $t_2$  (Fig. 39.3). If in the system  $S$  the coordinates of the particle are  $A_1 (x_1, y_1, z_1)$ , we obtain for the system  $S'$  from (39.1)

$$x'_1 = x_1 - vt_1, \quad y'_1 = y_1, \quad z'_1 = z_1$$

By analogy, for point  $A_2$  we obtain

$$x'_2 = x_2 - vt_2, \quad y'_2 = y_2, \quad z'_2 = z_2$$

Find the difference between the values  $x'_2$  and  $x'_1$ :

$$x'_2 - x'_1 = x_2 - x_1 - v(t_2 - t_1)$$

or

$$\frac{x'_2 - x'_1}{t_2 - t_1} = \frac{x_2 - x_1}{t_2 - t_1} - v$$

Since

$$\frac{x'_2 - x'_1}{t_2 - t_1} = u' \quad \text{and} \quad \frac{x_2 - x_1}{t_2 - t_1} = u$$

it follows that

$$u' = u - v, \quad \text{or} \quad u = v + u' \quad (39.3)$$

It can easily be seen that if a particle moves in the primed system in an arbitrary direction at a constant speed  $u'$  and the primed system itself in an arbitrary direction at a speed  $v$  with respect to the unprimed, we obtain for the components of velocities in the direction  $X$

$$u_x = v_x + u'_x$$

Similar formulae are obtained for the velocity components along the other coordinate axes:

$$u_y = v_y + u'_y, \quad u_z = v_z + u'_z$$

This means that the vector  $\mathbf{u}$  is equal to the sum of vectors  $\mathbf{v}$  and  $\mathbf{u}'$ :

$$\mathbf{u} = \mathbf{v} + \mathbf{u}' \quad (39.4)$$

Formula (39.4) expresses the *velocity-composition law of classical mechanics*.

Let us return to our systems  $S$  and  $S'$  and consider the case when a particle moves along the axes  $X$  and  $X'$  with acceleration.\* In this case, if at some moment of time  $u = u' + v$ , then after a time interval  $\Delta t$  the velocities of the particle in both systems will be increased and we shall obtain new velocities

$$(u + \Delta u) = (u' + \Delta u') + v \quad (39.5)$$

since the primed system continues to move with a constant velocity  $v$  with respect to the unprimed.

Subtracting  $u = u' + v$  from (39.5), we obtain

$$\Delta u = \Delta u' \quad (39.5a)$$

that is, the increase in the velocities are the same in both systems. Dividing (39.5a) by  $\Delta t$  we obtain

$$\frac{\Delta u}{\Delta t} = \frac{\Delta u'}{\Delta t}, \quad \text{or} \quad a = a' \quad (39.6)$$

\* For motion in a straight line the formulae can be written in scalar form.

Multiplying both sides of this equation by the particle's mass, which in classical mechanics is the same in all inertial frames, we obtain

$$ma = ma' \quad (39.7)$$

Since force is the product of the mass and the acceleration which the force imparts to the body, it follows from (39.7) that

$$F = F' \quad (39.7a)$$

(If several forces act on the body,  $F$  and  $F'$  should be taken to mean the resultant forces.)

We have established that the accelerations and the forces remain the same in all inertial frames, that is, they do not change when one frame is substituted for another.

This proposition is formulated in mechanics as follows: the equations expressing Newton's laws are invariant with respect to Galilean transformations. Another way to express it is that if Newton's laws are valid for one inertial frame, they must be valid for any other inertial frame. This is a corollary of Galileo's principle of relativity.

Note that this statement is the essence of the law of inertia, which states that uniform rectilinear motion takes place in the absence of forces. But if bodies in this frame interact, their resulting motion will be such as it would have been if there had been no common motion of the frame. For instance, observing mechanical motion on the Earth we detect nothing which would point to the motion of the Earth itself, travelling at a speed of 30 km/s.

### 39-3 Experimental Foundations of Einstein's Special Theory of Relativity

The Galilean transformations, together with the invariance of Newton's laws, are valid only if the course of time is identical in all inertial frames, that is, if any two events are simultaneous in one frame, they are also simultaneous in all other frames. Next we assume the length of a segment  $AB$  and the mass of a body to be identical in all inertial reference frames.

The point about an identical course of time in all reference frames was regarded as obvious in classical mechanics. Newton introduced into classical mechanics the postulates of absolute time and absolute space. "Absolute space", wrote Newton, "in its own nature, without relation to any-



thing external, remains similar and immovable.”\* He continued that the true course of time is not liable to change. In Newton’s opinion the course of time had nothing to do with the reference frame and was an absolute category.

As we have already remarked, the reference frame of the Earth cannot always be regarded as inertial. Already in the Copernican system of the Universe it was assumed that one used a reference frame stationary in space as a true inertial reference frame, not the Earth’s frame.

Newton’s postulate of absolute space implies an absolutely stationary reference frame. He assumed that out of a multitude of inertial reference frames moving with respect to one another and any one of which, as we know, can be chosen as a stationary frame there is one privileged frame fixed in absolute space which is really stationary. The motion of all bodies with respect to it is really true or absolute.

There are no experiments which could possibly prove the motion of inertial frames in Newtonian absolute space. Observing from an inertial frame the motion of all the bodies in the Universe which move independently of our frame, the only conclusion we can draw is that we move with respect to those bodies, but not in an absolute sense. A space devoid of all matter would be totally unobservable.

Since it is impossible to detect the motion of an inertial frame with the aid of mechanical methods, it may still be possible to do so by optical methods. Attempts were made at the end of the last century.

Since the Earth moves in an orbit in universal space (which was presumed to be absolutely immobile and the velocity of light in it to be the same in all directions and equal to  $c$ ), its motion should affect the velocity of light on the ground. There should be a difference between the velocity of light in the direction of the orbital motion of the Earth and in the direction normal to it.

Michelson and the American chemist Edward W. Morley (1838-1923) compared the velocities of light in these directions using the interference method. However, they failed to detect any effect of the Earth’s motion on the velocity of light. The experiments were repeated many times, but the velocity of light in the Earth’s reference frame turned out to be identical in all directions. This means that the motion of the Earth does not in any way affect the velocity

\* *Sir Isaac Newton’s Mathematical Principles of Natural Philosophy and His System of the World*, Andrew Motte, trans., University of California Press, Berkeley, 1962, Vol. 1, p. 6.

of light and that the velocity-composition law accepted in classical mechanics does not hold in this case.

Doubts then appeared as to whether the mass of a body always remains constant. The ratio  $e/m$  for electrons in cathode rays ( $e$  being the electron charge and  $m$  the electron mass) turned out to diminish at high electron speeds with the increase in speed. From the point of view of Newtonian mechanics this was incomprehensible, since  $e$  and  $m$  should stay constant.

To explain all these contradictions a new theory was required, based on assumptions different from those accepted in Newtonian mechanics. This theory was evolved by Einstein in the beginning of this century on the basis of new postulates, which were in full agreement with the Michelson-Morley experiment and with all other experiments.

The above should not be taken to mean that Newtonian mechanics is incorrect. It is in contradiction only with the experiments involving the velocity of light or the motion of particles at speeds approaching the speed of light  $c$ . In all other cases, when we deal with speeds of motion much less than the speed of light, classical mechanics is confirmed by experiment. This means that in developing the new mechanics one should adhere to the *principle of correspondence*, that is, one should include classical Newtonian mechanics in the new mechanics as a limiting case.

Hence, the laws of new mechanics become the same as Newton's laws at speeds of motion  $v$  small as compared with the speed of light  $c$ . The new mechanics was termed *relativistic mechanics*. Hence, relativistic mechanics does not supersede classical mechanics but simply establishes limits to its validity.

Let us now turn to Einstein's postulates.

(1) *The principle of constant light velocity*: the velocity of light in a vacuum ( $c$ ) is identical in all inertial frames and is independent of direction. It is independent of the motion of source or observer.

(2) *The principle of relativity*: there are no physical experiments (mechanical, electrical, optical) which when carried out in some inertial reference frame, could establish whether the system is at rest or in a state of rectilinear uniform motion. The physical laws are absolutely identical in all inertial systems.

Thus, Einstein's second postulate generalizes Galileo's relativity principle formulated for mechanical phenomena to include all phenomena in nature. Einstein's relativity principle establishes the complete equality of all inertial

reference frames and rejects Newton's idea of absolute space.

The term for Einstein's theory describing a phenomena taking place in inertial reference frames on the basis of the postulates cited above is *special theory of relativity*. The fundamentals of this theory will now be discussed.

The special theory of relativity had to drop the notions of space and time accepted in classical mechanics and customary to our mind because they disagreed with the principle of the constant velocity of light which had been established experimentally.

Not only absolute space, whose properties are independent of the reference frame and matter, became meaningless but absolute time as well. Time, too, turned out to be relative in the sense that definite moments of time and time intervals could be spoken of only in conjunction with some reference frame. Next it was established that the dimensions of bodies found from experiment are also relative and have to be related to a definite reference frame.

### **39-4 What Are Simultaneous Events in Special Relativity?**

Consider now the problem of synchronizing clocks and of events simultaneous in different reference frames, taking account of Einstein's postulates.

In Newtonian mechanics the true (standard) course of absolute time is not liable to change and does not depend on whether the motions are fast, slow or nonexistent. It was presumed that such concepts as moment of time, simultaneity, earlier, later have an inherent meaning equally valid throughout the Universe.

From the viewpoint of Einstein's theory of relativity there is no such concept as absolute simultaneity, just as there is no absolute time.

To establish whether two events at different points  $A$  and  $B$  take place simultaneously one must have accurate clocks at each point and be sure that the clocks are synchronized. To this end one can bring the clocks together at some point, adjust them (make them synchronous) and again bring them apart to points  $A$  and  $B$ . One can also make use of time signals to compare the readings of the clocks stationed at different points. Both methods are used in practice. Every ship, for instance, has a chronometer precisely synchronized by a standard clock in the

port of departure. In addition, radio-time signals are used to check it while at sea.

For checking clocks, electromagnetic signals (for instance, light) are especially convenient since their speed is the highest of all speeds of signals known to us. Still, this speed, although very great, is finite. Thus, for example, an inhabitant of Leningrad receives the time signal from Moscow 0.002 s late. This is not long, but it can be taken into account if the speed of electromagnetic signals and the distance from Moscow to Leningrad are known.

Einstein suggested a convenient method for checking the synchronism of clocks spaced far apart, a method which takes account of the time the signal travels from one clock to another. Note that the possibility of such a correction for the time of propagation of the signal follows from the first postulate of Einstein, concerning the constant speed of light.

First of all let us agree that we shall be able to place any number of synchronous clocks at a definite point in space. Leave one clock at  $A$  and take the others to the desired points  $B$ ,  $C$ , etc. (Fig. 39.4). To check whether the clocks,

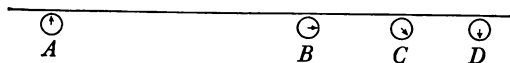


Fig. 39.4 Clock  $A$  indicates when light signal has been sent; all other clocks indicate when they receive light signal from  $A$ .

for instance at  $A$  and  $B$ , are still synchronous we adopt the following procedure: at the moment  $t_A$  (by the clock at  $A$ ) we send a light signal in the direction of  $B$ . Let it arrive at  $B$  at the moment  $t_B$  (by the clock at  $B$ ), be reflected by a mirror and return to  $A$  to be received there at the moment  $t'_A$  (by the clock at  $A$ ). According to Einstein, the clocks at  $A$  and  $B$  are synchronous if the readings of the clocks satisfy the relation

$$t_B - t_A = t'_A - t_B \quad (39.8)$$

For instance, let a signal from Moscow to Leningrad be sent at the moment  $t_A = 0$  (by the clock in Moscow) to be received at the moment  $t_B = 0.002$  s (by the clock in Leningrad) and reflected by a mirror back to Moscow to be registered there at the moment  $t'_A = 0.004$  s (by the clock in Moscow). In this case we can claim that the clocks in Moscow and Leningrad are synchronous,  $(0.002 - 0)c = (0.004 - 0.002)c$ .

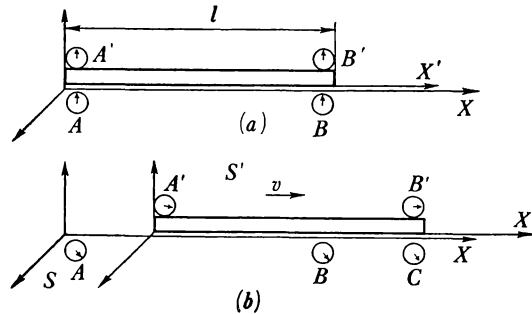
In this way one can check the running of the clocks at points  $B$ ,  $C$ , etc. against the clock at point  $A$  if all the

clocks are at rest with respect to one another. The farther the clocks are from  $A$  the more time the light signal will take to reach them from  $A$ . Therefore, if a signal from  $A$  is sent simultaneously to all the clocks, their readings at the time it is received at  $B, C$ , etc. will be different (see Fig. 39.4).

Now let us imagine a long rigid rod of length  $l_{AB}$  with its ends  $A'$  and  $B'$  carrying clocks at points  $A$  and  $B$  respectively (Fig. 39.5). When the rod is at rest in the reference frame  $S$  the clocks at  $A'$  and  $B'$  are synchronous with the clocks at  $A, B, C$ , etc.

Now let the rod move from  $A$  to  $B$  at a speed  $v$  with respect to the observer remaining at rest in the reference frame  $S$

Fig. 39.5 (a) Clock indicates when light signal has been sent from  $A$  to  $B$  and  $B'$ ; (b) clock indicates when signal reaches end of rod,  $B'$ .



at point  $A$ . Will the clocks at the ends of the moving rod  $A'$  and  $B'$  still be synchronous with the clocks at  $A, B$ , etc., which remain in the stationary reference frame  $S$ , if all the clocks continue to run with ideal precision?

Let a light signal be sent from  $A$  to  $B$  and  $B'$  at the moment the leading end of the rod  $A'$  coincides with the point  $A$  in the frame  $S$  (Fig. 39.5a). In the time light travels with a velocity  $c$  from the clock at  $A'$  to the clock at  $B'$  the rod will be displaced in the frame  $S$  and its trailing end  $B'$  will be opposite some point  $C$  (Fig. 39.5b). From the point of view of the observer on the rod the time the signal traveled from  $A'$  to  $B'$  is

$$t'_B - t'_A = l_{A'B'}/c \quad (39.9)$$

An observer in the frame  $S$  sees that to reach the end of the rod the light signal has to travel a distance  $l_{AC}$ . Hence, from his point of view the time the signal takes to reach the end of the rod from  $A$  is

$$t_C - t_A = l_{AC}/c \quad (39.10)$$

Since  $t_A = t'_A$  we obtain  $t'_B \neq t_C$  because  $l_{AC}$  is greater than  $l_{A'B'}$ . Hence, the times the signal travels

from the front end  $A'$  to the tail end of the rod  $B'$  as measured by the clock of the observer at rest in the frame  $S$  and by the clock of the observer at rest in the frame  $S'$  will be different. This means that an ideally precise clock in the frame  $S$  is not synchronous with an identical clock in the frame  $S'$ .

Thus, according to the theory of relativity each of the inertial frames moving with respect to each other has its own proper time measured by clocks at rest in this frame. Accordingly, events observed as simultaneous in one frame may appear as not simultaneous in another. In other words, there is no absolute simultaneity.

### 39-5 Lorentz Transformations

A mathematical analysis of phenomena taking place in inertial frames of reference carried out by Einstein on the basis of his postulates proved the Galilean transformations to be inconsistent with those postulates. Thus they had to be replaced by new ones, termed the *Lorentz transformations*.

To derive these transformations let us consider two inertial frames  $S$  and  $S'$  moving at a constant speed  $v$  with respect to each other. Choose in both frames orthogonal coordinate sets  $XYZ$  and  $X'Y'Z'$  such that the axes  $X$  and  $X'$  coincide with each other and with the direction of motion and the axes  $Y$  and  $Y'$ ,  $Z$  and  $Z'$  are parallel (Fig. 39.6a). Let the frame  $X'Y'Z'$  move with respect to the frame  $XYZ$  in the positive direction of  $X$ , that is, to the right in Fig. 39.6a. For the sake of simplicity we assume the origins  $O$  and  $O'$  to coincide at the initial moment. In that case the coordinate of point  $O'$  in the  $S$  frame will be  $vt$ , where  $t$  is the time measured in this frame.

Take a point  $A$  with an arbitrary coordinate  $x'$  in the frame  $S'$  on the  $X'$  axis and find the coordinate of the same point  $x$  in the  $S$  frame. The coordinate  $x'$  of this point in the  $S'$  frame is represented by the segment  $O'A$ . An observer at rest in the  $S$  frame records the positions of the ends of this segment at some moment  $t$  (by the clocks in his frame), the coordinates  $x$  (for point  $A$ ) and  $vt$  (for point  $O'$ ), and measures the length of this segment in the  $S$  frame as equal to  $(x - vt)$ .

In classical mechanics, as we know, the length of any segment is independent of the frame in which it is measured, and we could in such a case equate  $x'$  and  $(x - vt)$  to obtain Galilean transformations. From the point of view of the theory of relativity, however, one is not entitled

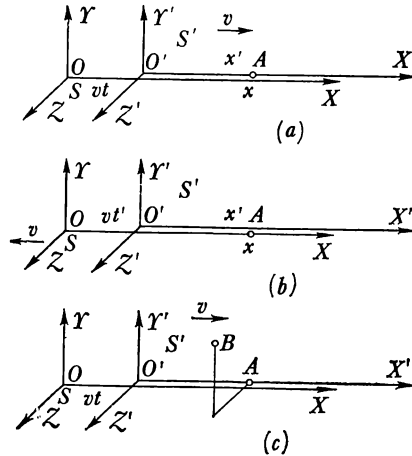
to assert that the lengths of the segment  $O'A$  measured in different reference frames must coincide. But they must at least be proportional because, if some segment is a number of times longer or shorter than another in the frame  $S'$ , it should be the same number of times longer or shorter in the frame  $S$ . Therefore we can write

$$x' = \alpha (x - vt) \quad (39.11)$$

where  $\alpha$  is, as yet, an unknown factor.

Have the frames  $S$  and  $S'$  change places. Now take a point  $A$  with an arbitrary coordinate  $x$  in the  $S$  frame.

**Fig. 39.6** Lorentz transformations. Frames  $S$  and  $S'$  move with respect to each other at speed  $v$ : (a) particle  $A$  is at rest in frame  $S'$ ; (b) particle  $B$  with arbitrary coordinates is at rest in frame  $S'$ .



The frame  $S$  may be regarded as one moving at a speed  $-v$  with respect to the  $S'$  frame (in the direction of negative  $X'$ , that is, to the left in Fig. 39.6b). The coordinate of the point  $O$  in the  $S'$  frame is  $-vt'$ , where  $t'$  is the time measured in the  $S'$  frame from the moment the origins  $O$  and  $O'$  coincided.

The length of the segment  $OA$  in the  $S$  frame is  $x$ . The length of this segment measured by the observer at rest in the  $S'$  frame is  $(x' + vt')$ , and we can again assume that

$$x = \alpha (x' + vt') \quad (39.12)$$

where  $\alpha$  is the same constant multiplier since the inertial frames  $S$  and  $S'$  are absolutely equivalent (in accordance with Einstein's second postulate).

The following method is used to find this factor. Let a light signal be sent along the axes  $X$  and  $X'$  from the common origin of the  $S$  and  $S'$  frames at the moment their

origins  $O$  and  $O'$  coincided. This signal will reach an arbitrary point  $x$  in the  $S$  frame and  $x'$  in the  $S'$  frame in time  $t$  measured in the  $S$  frame and in time  $t'$  measured in the  $S'$  frame. Then applying Einstein's first postulate ( $c = c'$ ) we can write  $x = ct$  and  $x' = ct'$ . Substituting these expressions for  $x$  and  $x'$  into (39.11) and (39.12), we obtain

$$ct' = \alpha(ct - vt), \quad ct = \alpha(ct' + vt')$$

Factoring out time and multiplying the relations we obtain

$$c^2 t' t = \alpha^2 t' t (c - v)(c + v), \quad \text{or} \quad c^2 = \alpha^2 (c^2 - v^2)$$

whence

$$\alpha = \frac{c}{\pm \sqrt{c^2 - v^2}} = \frac{1}{\pm \sqrt{1 - v^2/c^2}} \quad (39.13)$$

(Explain why the negative solution has no physical meaning.) Taking the positive value of the root we obtain finally

$$\alpha = \frac{1}{\sqrt{1 - v^2/c^2}} \quad (39.13a)$$

Substituting this value of  $\alpha$  into (39.11) and (39.12) we obtain the Lorentz transformations

$$x' = \frac{x - vt}{\sqrt{1 - v^2/c^2}} \quad (39.14)$$

$$x = \frac{x' + vt'}{\sqrt{1 - v^2/c^2}} \quad (39.15)$$

The Lorentz transformation formulae relating  $t$  and  $t'$  can be found by the following method. Divide the left-hand and the right-hand sides of (39.14) by  $c$ .

$$\frac{x'}{c} = \frac{x/c - vt/c}{\sqrt{1 - v^2/c^2}}, \quad \text{or} \quad t' = \frac{t - vx/c}{\sqrt{1 - v^2/c^2}}$$

Since  $t = x/c$ , the last formula can be rewritten in the form

$$t' = \frac{t - vx/c^2}{\sqrt{1 - v^2/c^2}} \quad (39.16)$$

In the same way we obtain from (39.15)

$$t = \frac{t' + vx'/c^2}{\sqrt{1 - v^2/c^2}} \quad (39.17)$$

If in the  $S'$  frame we were to take an arbitrary point  $B$  with the same coordinate  $x'$  as that of point  $A$  but not lying on the  $X'$  axis (Fig. 39.6c), we obtain the same relations for the coordinates  $x$  and  $x'$  of this point and for the times  $t$  and  $t'$  in our frames  $S$  and  $S'$ , while the other two coordi-



nates  $y'$  and  $z'$  of the point  $B$  in the  $S'$  system would, obviously, coincide with the respective  $y$  and  $z$  coordinates in the  $S$  frame.

For the sake of comparison we write Galilean transformations for one-dimensional motion in the  $OX$  direction alongside Lorentz transformations:

<i>Galilean transformations</i>	<i>Lorentz transformations</i>
<i>Direct</i>	<i>Direct</i>
$x' = x - vt$	$x' = \frac{x - vt}{\sqrt{1 - v^2/c^2}}$
$y' = y$	$y' = y$
$z' = z$	$z' = z$
$t' = t$	$t' = \frac{t - vx/c^2}{\sqrt{1 - v^2/c^2}}$
<i>Inverse</i>	<i>Inverse</i>
$x = x' + vt'$	$x = \frac{x' + vt'}{\sqrt{1 - v^2/c^2}}$
$y = y'$	$y = y'$
$z = z'$	$z = z'$
$t = t'$	$t = \frac{t' + vx'/c^2}{\sqrt{1 - v^2/c^2}}$

Note that in full accord with the principle of equivalence of inertial frames the formulae of the direct and inverse transformations are of the same form since, if the speed of the frame  $S'$  with respect to the frame  $S$  is  $v$ , the speed of the frame  $S$  with respect to the frame  $S'$  is  $-v$ .

Comparing the Galilean and Lorentz transformations, we see that, for speeds of motion of one inertial frame with respect to another small as compared with  $c$ , the Lorentz transformations are reduced to the Galilean, so that the principle of correspondence mentioned above is satisfied.

### 39-6 Length and Time Interval in Special Relativity

Consider the problem of measuring the length of a rod in a stationary and a moving reference frame.

If a rod is at rest with respect to the observer, its length can be measured by simply matching the beginning and

the end of the rod against a scale. The length measured in this way is termed *proper length* of a rod and is denoted by  $l_0$ . This is the length we normally obtain as a result of measurement of some linear dimension of a body during experiment.

Now imagine an observer at rest in an inertial frame  $S$  and a rod parallel to the  $X$  axis of this frame and moving along the  $X$  axis at a speed  $v$ . How can such an observer measure the length of a moving rod  $l$ ?

The usual method of measuring length is evidently useless in this case. A possible method is as follows: the observer

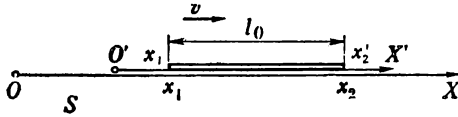


Fig. 39.7 Measuring length of moving rod.

at rest in the frame  $S$  at some moment of time measured by the clocks in his frame registers the positions of the front and the tail ends of the rod (Fig. 39.7) and then measures the distance  $l$  between the clocks that at that moment are opposite the front ( $x_1$ ) and the tail ( $x_2$ ) ends of the rod. This distance  $l = x_2 - x_1$  will be the length of a moving rod in the stationary reference frame  $S$ .

Now compare  $l$  with the proper length of the rod  $l_0$ . Let the coordinates of the ends of the rod in the  $S'$  frame (in which it is at rest) be  $x'_1$  and  $x'_2$ . Then from (39.18) we get

$$x'_1 = \frac{x_1 - vt_1}{\sqrt{1 - v^2/c^2}} \quad \text{and} \quad x'_2 = \frac{x_2 - vt_2}{\sqrt{1 - v^2/c^2}}$$

whence after subtraction we obtain

$$x'_2 - x'_1 = \frac{x_2 - x_1 - v(t_2 - t_1)}{\sqrt{1 - v^2/c^2}} \quad (39.20a)$$

Since the coordinates of the rod's ends  $x_1$  and  $x_2$  in the  $S$  frame are measured simultaneously from the point of view of the observer in the frame, it follows that  $t_2 = t_1$ . Therefore from (39.20a) we obtain

$$x'_2 - x'_1 = \frac{x_2 - x_1}{\sqrt{1 - v^2/c^2}} \quad (39.20)$$

Since  $x'_2 - x'_1$  is the distance between the front and the tail end of the rod in the  $S'$  frame in which it is at rest,

$x'_2 - x'_1 = l_0$ . But since  $x_2 - x_1 = l$ , we obtain\*

$$l_0 = \frac{l}{\sqrt{1 - v^2/c^2}} \quad (39.21)$$

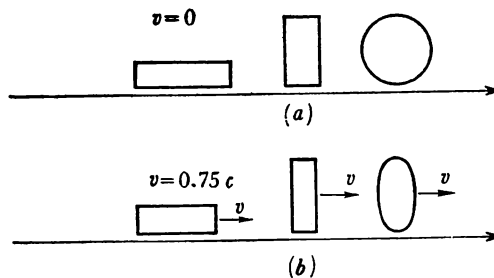
It follows from (39.21) that

$$l = l_0 \sqrt{1 - v^2/c^2} \quad (39.22)$$

Hence, we see that the results of measuring the length of a rod are relative and depend on its speed with respect to a reference frame: the length is always less than the proper length  $l_0$  (the factor  $\sqrt{1 - v^2/c^2}$  is less than unity), decreasing with the speed of motion of the rod with respect to the reference frame in which it is being measured.

However, if the rod is rotated by  $90^\circ$ , that is, placed at right angles to the  $X$  axis and to the direction of motion, its length will not change as compared with  $l_0$  since  $y' = y$  and  $z' = z$ . Thus, only the dimensions of a body in the direction of motion are contracted. For instance, measuring the dimensions of a moving sphere we shall find them to be curtailed in the direction of motion, that is, the sphere will appear to us as an ellipsoid of revolution (Fig. 39.8).

**Fig. 39.8** Contraction of size of moving bodies: (a) bodies at rest; (b) bodies moving at speed  $v$ .



Note that this effect is relative. For instance, if one meter-long ruler is at rest in the frame  $S$  and any other in the frame  $S'$  and the frames are moving with respect to one another at a speed  $v$ , then two observers, one of whom is at rest in frame  $S$  and the other in frame  $S'$ , will see the ruler moving with respect to him in a contracted form.

\* This result was to be expected since in deriving the Lorentz transformations and comparing the lengths of the segment  $O'A$  in the frames  $S'$  and  $S$  we found that they should be proportional, the proportionality factor being

$$\alpha = \frac{1}{\sqrt{1 - v^2/c^2}}.$$

Consider now the problem of the relativity of time intervals. We have already convinced ourselves that ideal identical clocks in two reference frames moving with respect to one another do not run synchronously. Imagine an observer at rest with respect to one clock and moving with respect to the other at a speed  $v$ . Is it possible to find out which clock runs faster and which slower? In other words, what time interval between two events measured by both clocks will be greater? The answer can be found with the aid of the Lorentz transformations.

Let one observer be in a railcar and have a clock at rest in the car. We denote the reference frame of the car by  $S'$ . Let another observer with his clock be at rest with respect to the Earth, and let the train be moving at a speed  $v$ . The notation for the Earth's reference frame will be  $S$ .

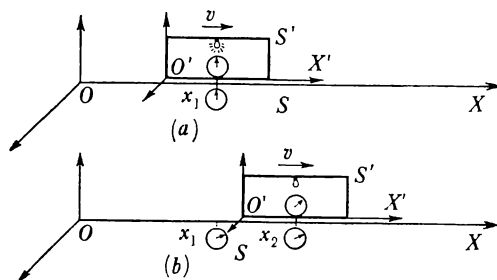


Fig. 39.9 Clocks in railcar and on platform indicate moments of time: (a) of the first event (lamp lit up in railcar); (b) of the second event (lamp went out).

Suppose now that at a moment  $t'_1$  (Fig. 39.9a) a lamp was lit in the railcar (a definite event took place) and at the next moment  $t'_2$  (Fig. 39.9b) it was put out (a new event took place). For the observer in the car both events took place at the same point in space (in the car) but at different moments of time  $t'_1$  and  $t'_2$ .

We shall apply the term *interval of time* to the difference between two definite moments of time and denote it by  $T$ . The term for the interval of time between two events measured in the reference frame in which these events took place is interval of *proper time*  $T_0$ . Hence, for the observer in the railcar we have

$$t'_2 - t'_1 = T_0$$

For the observer on the Earth both events took place at different points of space and at different moments of time,  $t_1$  and  $t_2$ , by his clock. Indeed, the lamp was lit in one place and put out in another since during the time it burned the train covered a definite distance with respect to the

Earth. For the observer on the ground this interval will be equal to

$$t_2 - t_1 = T$$

Compare now the interval  $T$  with the interval of proper time  $T_0$ . To this end we express  $t_1$  and  $t_2$  in terms of the coordinates of the primed frame

$$t_1 = \frac{t'_1 + vx'_1/c^2}{\sqrt{1-v^2/c^2}} \quad \text{and} \quad t_2 = \frac{t'_2 + vx'_2/c^2}{\sqrt{1-v^2/c^2}}$$

and find the difference  $t_2 - t_1$ :

$$t_2 - t_1 = \frac{(t'_2 - t'_1) + v(x'_2 - x'_1)/c^2}{\sqrt{1-v^2/c^2}} \quad (39.23)$$

Since in the primed frame both events took place at the same point of space,  $x'_2 = x'_1$  and  $x'_2 - x'_1 = 0$ . Therefore we obtain from (39.23)

$$t_2 - t_1 = \frac{t'_2 - t'_1}{\sqrt{1-v^2/c^2}} \quad (39.24)$$

or

$$T = \frac{T_0}{\sqrt{1-v^2/c^2}} \quad (39.25)$$

It follows from (39.25) that  $T_0 < T$ , that is, the interval of proper time is shorter. Hence, when observers at rest in different reference frames measure time intervals between two events, the clock in the frame in which the events take place is found to run slower.

If an observer on the platform watches the events taking place in a moving railcar, he gets the impression that the clock in the car runs slower than his own, that is, his clock measures a greater time between two events taking place in the car than the clock in the car. Conversely, if an observer is in a moving railcar and watches the events on the platform he gets the impression that the clock on the platform runs slower than the clock in the car, that is, the time interval between two events taking place on the platform measured by his clock is longer than one measured by the clock on the platform. From the point of view of each observer the clocks moving with respect to him slow down as compared with his clock.

Here the relative nature of time intervals is seen quite clearly, for each of the observers imagines that his counter-

part's clock lags behind his own. This result supports the equivalence of all inertial reference frames.

The dependence of time intervals on the reference frame can be seen in experiment.

Here is one example. The atmosphere of the Earth is constantly bombarded by cosmic rays consisting of particles moving at enormous speeds. When those particles collide in the upper layers of the atmosphere with atoms of atmospheric nitrogen or oxygen,  $\pi$ -mesons are formed (see Sections 40-10 and 41-1). They are unstable and live only a very short time (their lifetime is very short).

We can obtain  $\pi$ -mesons artificially in big accelerators. The average lifetime of those  $\pi$ -mesons, that is, the average time interval between their birth and decay, was measured in the laboratory. The velocity of these  $\pi$ -mesons is not large, far below  $c$ . Therefore their lifetime  $T_0$ , measured in experiment, can be regarded as the proper lifetime of the  $\pi$ -meson. It turns out to be quite small, of the order of several hundredths of a microsecond:  $T_0 = 2 \times 10^{-8}$  s. This means that even if a  $\pi$ -meson were to travel at a speed close to that of light, then from the classical point of view it would not be able to cover in its lifetime a distance greater than 6 m, since  $l = cT_0 = 3 \times 10^8 \text{ m/s} \times 2 \times 10^{-8} \text{ s} = 6 \text{ m}$ .

But  $\pi$ -mesons have been discovered at ground level, that is, they have penetrated the atmosphere and reached the ground, thus having covered a distance of the order of 30 km without decay.

This is explained by the slowing down of time: each  $\pi$ -meson can be imagined to carry its own clock which ticks off its proper lifetime  $T_0$ . However, to the observer on the ground the lifetime  $T$  of the  $\pi$ -meson appears much longer, in compliance with formula (39.25), since the  $\pi$ -meson's speed is actually very close to that of light.

This fact can also be interpreted in another way: for the  $\pi$ -meson moving at a speed close to  $c$ , the terrestrial measures of length appear greatly contracted in the direction of the  $\pi$ -meson's motion with respect to the Earth, in accordance with formula (39.22). In other words, if one takes into account the proper lifetime of the  $\pi$ -meson  $T_0$ , one must measure the terrestrial distances in the reference frame of the  $\pi$ -meson as well.

This example shows clearly that the concept of measurement has no absolute meaning and that numbers denoting distance and time are not absolute, being meaningful **only** for a definite reference frame.

### 39-7 The Relativistic Velocity-Composition Law

Another important corollary of the Lorentz transformations is a new velocity-composition law, which differs from the classical velocity-composition equation.

Let two inertial reference frames  $S$  and  $S'$  move with respect to each other at a speed  $v$ . As before, we can choose sets of coordinates  $XYZ$  for  $S$  and  $X'Y'Z'$  for  $S'$  so that their  $X$  and  $X'$  axes coincide with each other and with the direction of relative motion of the frames, and so that the axes  $Y$  and  $Z$  are parallel to the axes  $Y'$  and  $Z'$  respectively.

Now let a particle move in the  $S'$  frame along the  $X'$  axis. In this case, if the positions of the particle in the  $S'$  frame at the moments  $t'_1$  and  $t'_2$  are  $x'_1$  and  $x'_2$  respectively, then the speed of the particle in the  $S'$  frame will be  $u' = (x'_2 - x'_1)/(t'_2 - t'_1)$ . The speed of this particle measured in the  $S$  frame will be  $u = (x_2 - x_1)/(t_2 - t_1)$ . Making use of Lorentz transformations, we find the relation between the speeds  $v$ ,  $u'$  and  $u$ . To this end we express the primed quantities in terms of the unprimed:

$$x'_2 = \frac{x_2 - vt_2}{\sqrt{1 - v^2/c^2}} \quad \text{and} \quad x'_1 = \frac{x_1 - vt_1}{\sqrt{1 - v^2/c^2}}$$

Subtracting the second equation from the first, we obtain

$$x'_2 - x'_1 = \frac{x_2 - x_1 - v(t_2 - t_1)}{\sqrt{1 - v^2/c^2}} \quad (39.26)$$

Also

$$t'_2 = \frac{t_2 - vx_2/c^2}{\sqrt{1 - v^2/c^2}} \quad \text{and} \quad t'_1 = \frac{t_1 - vx_1/c^2}{\sqrt{1 - v^2/c^2}}$$

Subtracting the second equation from the first, we obtain

$$t'_2 - t'_1 = \frac{t_2 - t_1 - v(x_2 - x_1)/c^2}{\sqrt{1 - v^2/c^2}} \quad (39.27)$$

Dividing (39.26) by (39.27), we obtain

$$\frac{x'_2 - x'_1}{t'_2 - t'_1} = \frac{(x_2 - x_1) - v(t_2 - t_1)}{(t_2 - t_1) - v(x_2 - x_1)/c^2}$$

If we then divide both the numerator and the denominator of the right-hand side by  $(t_2 - t_1)$ , we obtain

$$\frac{x'_2 - x'_1}{t'_2 - t'_1} = \frac{\frac{x_2 - x_1}{t_2 - t_1} - v}{1 - \frac{v}{c^2} \frac{x_2 - x_1}{t_2 - t_1}}$$

But since  $\frac{x'_2 - x'_1}{t'_2 - t'_1} = u'$  and  $\frac{x_2 - x_1}{t_2 - t_1} = u$ , we have

$$u' = \frac{u - v}{1 - vu/c^2} \quad (39.28)$$

Making use of the inverse Lorentz transformations from the parameters of the  $S'$  frame to the parameters of the  $S$  frame, we obtain a similar formula in which  $-v$  is substituted for  $v$ :

$$u = \frac{u' + v}{1 + vu'/c^2} \quad (39.29)$$

This is in full accord with the principle of equivalence of all inertial reference frames.

It can easily be seen that for speeds  $v$ ,  $u'$  and  $u$ , small in comparison with the speed of light  $c$ , the formulae (39.28) and (39.29) of the relativistic velocity-composition law reduce to the corresponding formulae of classical mechanics.

Consider now another limiting case. Suppose we are dealing with a ray of light propagating in the  $S'$  frame along the  $X'$  axis. Then  $u' = c$ , and we obtain for the velocity of propagation  $u$  of this ray of light in the  $S$  frame from (39.29)

$$u = \frac{c + v}{1 + vc/c^2} = c$$

This result agrees with Einstein's first postulate concerning the constant velocity of light.

Note that  $u$  will be equal to  $c$  even if the speed of relative motion of the frames is itself close to  $c$  (i.e. for  $v = c$ ). This supports the proposition that in the theory of relativity the summation of speeds can never produce a result exceeding the speed of light in a vacuum,  $c$ .

It should be remarked that it is the speed of light in a vacuum which is the limiting speed that cannot be exceeded. The speed of light in some medium equal to  $c/n$  (where  $n$  is the absolute refractive index of the medium) is not a limiting parameter; electrons in water can move at higher speeds than the velocity of light propagation in water (see Section 40-5).

### 39-8 Mass and Energy in Special Relativity

It was established as the result of studies of the  $\beta$ -radiation of radioactive substances that it consists of electrons moving at speeds close to the speed of light. Why is it



not possible, say by applying a very high electric field, to make them move at speeds exceeding that of light?

The answer is that, according to Einstein's theory, the mass of a definite body is a relative quantity. Its magnitude depends on the reference frame in which it is being measured, or, if measured in the same reference frame, on the speed of the moving body. The mass depends only on the magnitude of the speed and is independent of its direction. This dependence is expressed by the following formula:

$$m = \frac{m_0}{\sqrt{1 - v^2/c^2}} \quad (39.30)$$

Here  $m_0$  is the *rest mass*, that is, the mass of the body measured in a frame in which it is at rest. As long as the speed of motion is small in comparison with  $c$  the mass may be regarded as a constant, that is, independent of speed, as is done in classical mechanics. As the speed of the body approaches the speed of light the magnitude of its mass  $m$ , in compliance with formula (39.30), continuously increases and a steadily increasing force is needed to impart a constant speed increment to the body. The closer the speed of the body to the speed of light the more difficult it is to increase it further. At  $v = c$  the mass becomes infinite. Hence, it is impossible to make a body move at the speed of light.

Experiments involving the deflection of cathode rays in electric and magnetic fields have proved with great accuracy that the mass of electrons does indeed rise with speed in accordance with formula (39.30). It follows from formula (39.30) that the momentum of a body in the theory of relativity

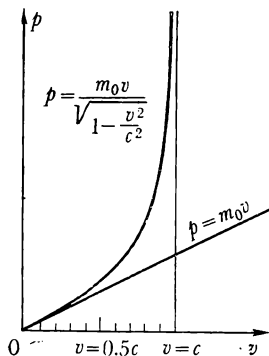
$$p = mv = \frac{m_0 v}{\sqrt{1 - v^2/c^2}} \quad (39.31)$$

is not proportional to its speed, as in the case in classical mechanics where the mass of a body is regarded as a constant. The momentum versus speed curve (Fig. 39.10) for speeds of motion small in comparison with  $c$  coincides with the straight line  $p = m_0 v$ , which is the momentum versus speed dependence of classical mechanics. But at speeds close to  $c$  these curves differ drastically.

The law of the conservation of momentum in a system remains valid in relativistic mechanics as well.

Formula (39.30) for the dependence of the mass on speed, presented here without derivation, can be derived from the momentum-conservation law with the aid of Einstein's velocity-composition law.

Fig. 39.10 Dependence of momentum of body on its speed: straight line, in classical mechanics; curved line, in theory of relativity.



### 39-9 Einstein's Mass-Energy Formula

The conclusion can be drawn from the contents of the preceding section that a body gaining kinetic energy increases its mass. It appears that a definite mass corresponds to a definite kinetic energy. Is this relationship true for other kinds of energy as well?

It turns out that a definite mass corresponds to energy of any kind. Thus, heating a body increases its mass a little. Radiation emitted by the Sun contains energy and by strength of this has mass; the Sun and the stars lose mass in the process of radiation. The term for the relation between mass and total energy is *Einstein's mass-energy formula*:

$$E = mc^2 \quad (39.32)$$

It follows that the total energy of a body is proportional to its mass. The masses of all bodies increase with the increase in their energies. In general, any variation in the energy (of a particle, body, system of bodies) in any form by an amount  $\Delta E$  entails a proportional variation in mass by an amount  $\Delta m$  in accordance with the formula

$$\Delta E = c^2 \Delta m \quad (39.33)$$

Thus, if a body loses energy  $\Delta E$  through the radiation of electromagnetic waves its mass diminishes by  $\Delta m$ , and if a body receives radiation it gains energy and its mass increases. In other words, according to the special theory of relativity, energy is proportional to mass.

The principle of proportionality between mass and energy established by the theory of relativity plays an enormous role in science, especially in atomic and nuclear physics. Formula (39.33) has been tested by experiment on numerous occasions and was brilliantly confirmed. Below we shall see how this formula is used to calculate the energy balance in nuclear reactions and to explain the phenomenon of the *annihilation* of particles, that is, their transformation into photons.

It would be interesting to compute, using formula (39.33), how the mass of the Sun changes with time. To calculate the amount of energy radiated by the Sun into space over a period of 1 s one computes the area of a sphere of a radius of 150 million kilometres, equal to the distance between the Sun and the Earth, with the Sun at its centre, and multiply this area by the solar constant, i.e. by the energy of the solar radiation passing through 1 m<sup>2</sup> of this surface per second (see Section 38-3). We obtain an enormous amount of

energy radiated by the Sun every second:  $3.8 \times 10^{26}$  J. The variation in mass corresponding to this radiation is, according to formula (39.32),  $4 \times 10^9$  kg. Hence, the mass of the Sun diminishes by 4 000 000 tonnes every second.

Find the expression for the kinetic energy of a body. Write Einstein's mass-energy formula (39.32) applicable to a body at rest in a selected reference frame

$$E_0 = m_0 c^2 \quad (39.34)$$

This relation shows that the body possesses a sort of latent energy, or *rest energy*, which always remains with it as long as the body continues to exist. Electrons and atoms are examples of gigantic energy concentrations. When a body (or a particle) for some reason or other ceases to exist, the energy  $E_0$  contained in it is simultaneously liberated. This energy (and mass equivalent to it) is gained by other bodies or particles taking part in the phenomenon (including particles of radiation—photons).

Now imagine the body to be set in motion at a speed  $v$ . In accordance with (39.30) the mass of the body increases and the total energy of the moving body  $E = mc^2$  exceeds its rest energy  $E_0$ . The difference between the total energy of the moving body and its rest energy is the kinetic energy:

$$K = E - E_0 = c^2 m - c^2 m_0 = c^2 (m - m_0),$$

$$\text{or} \quad K = c^2 \Delta m$$

For small values of  $v$  as compared with  $c$  this expression for  $K$  reduces to the classical one. To demonstrate it make use of (39.30):

$$\begin{aligned} K &= c^2 (m - m_0) = c^2 \left[ \frac{m_0}{\sqrt{1 - v^2/c^2}} - m_0 \right] \\ &= c^2 m_0 [(1 - v^2/c^2)^{-1/2} - 1] \end{aligned}$$

Apply the binomial expansion

$$(1 + a)^n = 1 + na + \frac{n(n-1)}{1 \times 2} a^2 + \dots$$

In our case  $n = -1/2$ , while  $a = -v^2/c^2$  is a small quantity. Therefore the third term containing  $a^2$  and the remaining terms are negligible and we may apply the approximate formula:

$$\left(1 - \frac{v^2}{c^2}\right)^{-1/2} \approx 1 + \left(-\frac{1}{2}\right) \left(-\frac{v^2}{c^2}\right) = 1 + \frac{1}{2} \frac{v^2}{c^2}$$

Hence

$$K = c^2 m_0 \left[ \left( 1 + \frac{1}{2} \frac{v^2}{c^2} \right) - 1 \right] = c^2 m_0 \frac{1}{2} \frac{v^2}{c^2} = \frac{m_0 v^2}{2}$$

This coincides with the well-known expression for the kinetic energy in classical mechanics.

### 39-10 Relation Between Momentum and Energy in Special Relativity

Let us find the relation between the momentum of a body or a particle and its energy.

Since  $E = mc^2$  and  $m = m_0 (1 - v^2/c^2)^{-1/2}$ , we can write

$$E = \frac{m_0 c^2}{\sqrt{1 - v^2/c^2}}$$

We square this equation and get

$$E^2 = \frac{m_0^2 c^4}{1 - v^2/c^2}, \quad E^2 - \frac{E^2 v^2}{c^2} = m_0^2 c^4$$

or

$$E^2 = m_0^2 c^4 + E^2 v^2 / c^2$$

Substituting  $mc^2$  for  $E$  in the right-hand side of the equation, we obtain

$$E^2 = m_0^2 c^4 + c^2 (mv)^2$$

Denoting the momentum  $mv$  by  $p$ , we obtain

$$E^2 = m_0^2 c^4 + c^2 p^2 \quad (39.35)$$

This relation establishes the connection between  $E$  and  $p$ :

$$E = \sqrt{m_0^2 c^4 + c^2 p^2}, \quad \text{or} \quad c^2 p^2 = E^2 - m_0^2 c^4 \quad (39.36)$$

$$p = \sqrt{\frac{E^2 - m_0^2 c^4}{c^2}} = \frac{\sqrt{E^2 - E_0^2}}{c} \quad (39.37)$$

We know that light is a flux of photons moving with a velocity  $c$ . Therefore the photon's momentum is  $p_{\text{ph}} = mc$ , where  $m$  is the "mass" of the moving photon. The photon's energy is expressed in terms of the momentum as follows:

$$E_{\text{ph}} = mc^2 = p_{\text{ph}} c$$

Hence  $p_{\text{ph}} = E_{\text{ph}}/c$ . From the latter expression we obtain, using formula (39.35),

$$E_{\text{ph}}^2 = m_0^2 c^4 + c^2 (E_{\text{ph}}/c)^2$$

whence

$$m_0^2 c^4 = 0, \quad \text{or} \quad m_0 = 0$$

that is, a photon has no rest mass.

According to Planck's formula (34.1) the photon's energy is  $E_{\text{ph}} = hf$ . Therefore

$$p_{\text{ph}} = E_{\text{ph}}/c = hf/c \quad (39.38)$$

Since  $f\lambda = c$ , we obtain

$$p_{\text{ph}} = h/\lambda \quad (39.39)$$

Hence, a light wave of a frequency  $f$  can be represented as a beam of particles (photons) having zero rest masses, energies equal to  $E_{\text{ph}} = hf$  and moments equal to  $p_{\text{ph}} = hf/c$  or  $p_{\text{ph}} = h/\lambda$ .

Photons can experience transformations into other particles, the energy and the momentum conservation laws being observed in the process.

part five

# **Nuclear Physics**

**40-1 Methods of Particle Detection**

In the first decades of the twentieth century research methods and apparatus were developed for atomic physics that enabled not only the fundamental problems of atomic structure to be solved, but the transmutations of chemical elements to be observed as well.

The main problem was to design an instrument that would be capable of registering a single charged particle used in the experiments—an ionized atom of some element or even an electron—or of making its track visible.

One of the first and simplest instruments used for registering particles was a screen coated with a luminescent compound. On such a screen a scintillation flash (from the Latin *scintillare* for to spark) appears at the spot hit by a particle of sufficient energy.

The first major instrument for registering particles was invented in 1908 by the German physicist Hans Geiger (1882-1945). Modified by his colleague W. Müller, the instrument was able to count the number of particles striking it. The operation of the *Geiger-Müller counter* is based on the ionization of gas atoms by charged particles flying through the gas.

The counter is made of a hollow metal cylinder of about three centimetres in diameter (Fig. 40.1) with a thin glass or aluminium window. Inside the cylinder there is a thin axi-

al metal filament insulated from it. The cylinder is filled with a rarefied gas, for instance argon. A voltage of about 1500 V, too low to cause a self-maintained discharge, is applied between the cylinder and the filament. The filament is grounded via a high resistor  $R$ . When a high-energy particle enters the chamber, it ionizes gas atoms along its

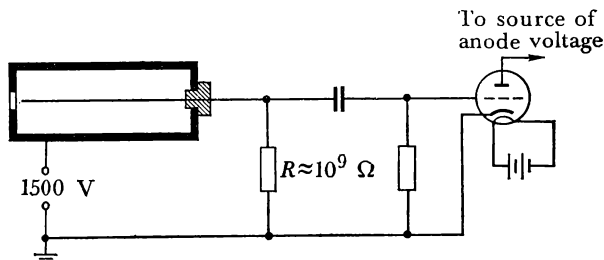


Fig. 40.1 Schematic representation of Geiger-Müller counter.

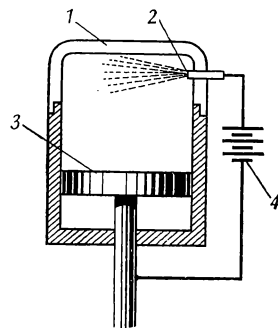
track and initiates a discharge between the walls and the filament. The discharge current creates a large voltage drop across the resistor  $R$ , and the voltage between the walls and the filament drops drastically. This instantly stops the discharge. After the current ceases to flow the voltage between the walls and the filament returns to its original value and the counter is ready to register a new particle. The voltage from the resistor  $R$  is applied to the input of an amplifier tube with a counting mechanism in its anode circuit.

The ability of high-energy particles to ionize gas atoms is also utilized in one of the most remarkable instruments of modern physics, the *Wilson cloud chamber*. In 1911 the Scottish physicist Charles Thomas R. Wilson (1869-1961) built an instrument which could be used to observe and photograph the tracks of charged particles.

The Wilson cloud chamber (Fig. 40.2) consists of a cylinder with a piston; the upper part of the cylinder is made of a transparent material. A small amount of water or alcohol is introduced into the cylinder. A mixture of vapour and air is thus formed in the cylinder. When the piston is depressed quickly, the mixture expands adiabatically and cools down. Consequently the air in the chamber becomes supersaturated with vapour.

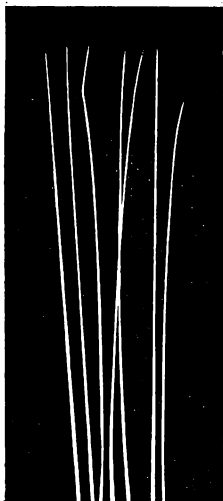
If the air is free of dust particles, the condensation of the excess vapour into liquid is made difficult by the absence of condensation centres. However, ions too can serve as condensation centres. Accordingly, if at this moment a charged particle is flying through the chamber leaving air ions in its

Fig. 40.2 Wilson cloud chamber: 1, transparent part of chamber; 2, grain of uranium salt; 3, piston; 4, battery.





**Fig. 40.3** Photograph of  $\alpha$ -particle tracks in Wilson cloud chamber.



**Fig. 40.4** Particle tracks in photoemulsion.



wake, vapour condenses on the chain of ions and the particle's track becomes marked by a thread of fog or becomes visible. The thermal motion of air soon disperses the fog threads; the tracks of particles are clearly visible for only about 0.1 s. This, however, is long enough to take photographs.

The shape of the track on the photograph often enables conclusions to be drawn as to the type of particle and its energy. For instance,  $\alpha$ -particles leave a comparatively thick solid track, protons a thinner track, and electrons a dotted one. A photograph of  $\alpha$ -particles in the Wilson cloud chamber is presented in Fig. 40.3.

To prepare the chamber for work and evacuate the remaining ions from it an electric field is set up inside the chamber, attracting the ions to the electrodes on which they are neutralized.

As was stated above, the condensation of supersaturated vapour, that is, the transformation of vapour into liquid, is used in the Wilson cloud chamber to obtain particle tracks. The reverse process, that is, the transformation of liquid into vapour, can also be used for this purpose. By first compressing a liquid in a closed space with the aid of a piston and then by rapidly withdrawing the piston the liquid, at an appropriate temperature, can be made superheated. When a charged particle flies through such a liquid the liquid boils in its wake, because the ions formed in it serve as vapourization centres. The result is that the track of the particles is marked by a chain of bubbles, that is, becomes visible. This is the principle behind operation of the bubble chamber.

The *bubble chamber* is better adapted for the study of high-energy particles than the Wilson cloud chamber because a particle moving in a liquid loses much more of its energy than in a gas. In many cases this fact makes it possible to establish the direction of motion of a particle and its energy with a much greater accuracy. At present there are bubble chambers of about 2 m in diameter in service. They are filled with liquid hydrogen. The particle tracks in liquid hydrogen are very distinct.

Another method used to register particles and obtain their tracks is the nuclear emulsion counter. Its operation is based on the fact that particles flying through the photoemulsion act on the particles of silver bromide, with the result that the particle's track is made visible by subsequent development (Fig. 40.4) and can be investigated under a microscope. To increase the length of the track thick emulsion layers are used.

## 40-2 Radioactivity

In 1896, the French physicist Antoine Henri Becquerel (1852-1908) discovered with the aid of a photographic plate that one of the uranium salts was a source of mysterious radiation. Becquerel established that this radiation is emitted by all uranium compounds including metallic uranium itself, or that uranium atoms are the source of radiation.

It is an experimental fact that uranium emits radiation continuously and no external action (temperature, pressure, etc.) has any effect on it, that is, uranium atoms emit radiation spontaneously. Uranium radiation was given the name radioactive radiation and the phenomenon itself was termed radioactivity.

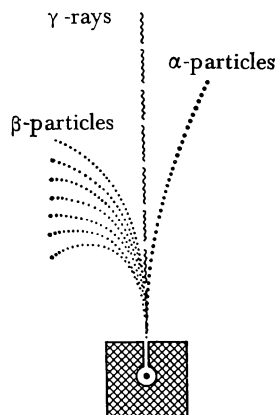
It was established as the result of research carried out by Becquerel, Rutherford, Pierre and Marie Curie and by other scientists that radioactive radiation has a complex composition and in a magnetic field is separated into three components (Fig. 40.5) termed  $\alpha$ -,  $\beta$ - and  $\gamma$ -rays. It was established that the  $\alpha$ -rays are, in effect, a flux of positively charged particles ( $\alpha$ -particles);  $\beta$ -rays a flux of fast electrons ( $\beta$ -particles); while  $\gamma$ -rays, which are not deflected in magnetic fields, are very short electromagnetic waves.

Some other heavy elements at the end of the Mendeleev Periodic Table were also found to exhibit radioactivity. In 1898, Pierre Curie (1859-1906) and his wife Marie (1867-1934) discovered the radioactivity of thorium and in the same year discovered two new chemical elements contained in uranium ore which, too, were radioactive. One of them, whose radioactivity proved to be about a million times that of uranium, was termed *radium* and the other *polonium*. In 1908 Rutherford, with the aid of spectroscopic analysis, discovered a radioactive gas, *radon*.

The discovery of radioactivity posed a problem for physicists: what is the origin of radioactive radiation? Most mysterious was its spontaneity. In 1903 Rutherford and the British scientist Frederick Soddy (1877-1956) offered the hypothesis that radioactive radiation is the result of spontaneous atomic decay. According to this hypothesis, the atoms of radioactive elements, in contrast to normal elements, are unstable and from time to time one or the other such atoms spontaneously decays. Subsequent research proved this hypothesis to have been true.

When the atomic structure had been established, it became clear that radioactive radiation is the result of the decay of the nuclei of atoms of radioactive elements, since only nuclei can be the source of positively charged  $\alpha$ -

Fig. 40.5 Radiation emitted by uranium ore separates in magnetic field into  $\alpha$ -,  $\beta$ - and  $\gamma$ -radiation (magnetic field points at reader).



particles. It was subsequently established that  $\beta$ -particles are also the result of nuclear decay.

The nature of  $\alpha$ -particles was finally established in 1908. The results of numerous experiments demonstrated that  $\alpha$ -particles are doubly ionized helium atoms, that is, helium nuclei. Rutherford undertook one experiment aimed directly at the matter:  $\alpha$ -particles were made to enter an evacuated vessel for a period of several days through a very thin window. A subsequent spectral analysis established the presence of helium in the vessel.

If a small amount of radon is placed inside a sealed ampoule, the intensity of its radioactivity will diminish with time. The explanation is that as the nuclei of the radon atoms decay the number of intact radon nuclei, that is, the amount of radioactive substance remaining in the ampoule, diminishes. Obviously, the more rapid the nuclear decay the sooner the radiation intensity diminishes. Different radioactive elements have different decay rates. Further, some elements include several radioactive isotopes with different decay rates.

The term for the quantity characterizing the decay rate of a radioactive isotope is *half-life* and its designation is  $T$ . The measure for a half-life is the time during which the number of atoms of a radioactive isotope diminished by half. For example, the half-life of radium is 1602 years. Accordingly, if one takes, say, one gram of radium, after 1602 years a half of it will remain (0.5 g), in 3204 the remaining amount will be a quarter (0.25 g), and so on.

The half-life of uranium is  $4.51 \times 10^9$  years, but that of radon only 3.82 days. The nuclei of some radioactive elements are so unstable that their half-life is measured in microseconds.

### 40-3 Transmutation of Elements

The decay of the atomic nuclei of a radioactive isotope of an element results in the formation of nuclei of the isotopes of other elements; for instance, radon and helium are produced as the result of the decay of radium. Thus, radioactive decay entails the transformation of the chemical element into another.

The chemical nature of atoms is known to be determined by their nuclei. In order to transform an atom of one chemical element into an atom of another element its atomic number  $Z$  has to be changed. Thus, the emission of an  $\alpha$ -particle decreases the atomic number  $Z$  by two units, it be-

coming equal to  $Z - 2$ ; the emission of a  $\beta$ -particle increases the atomic number by one and it becomes  $Z + 1$ . In this way the emission of  $\alpha$ -particles by the nuclei of radium atoms turns them into radon nuclei; the emission of  $\beta$ -particles by actinium nuclei makes thorium nuclei out of them. Such transformation can be expressed with the aid of the *radioactive displacement law*: a chemical element emitting an  $\alpha$ -particle is displaced by two positions to the left in the Mendeleev Periodic Table, and an element emitting a  $\beta$ -particle is displaced by one position to the right.

As for  $\gamma$ -radiation, it usually accompanies  $\alpha$ - or  $\beta$ -radiation. After emitting  $\alpha$ - or  $\beta$ -particles the atomic nucleus often finds itself in an excited state, that is, it has an excess energy, and in going over to a lower energy level (to the ground state) emits a  $\gamma$ -quantum.

New nuclei produced as the result of radioactive decay may also be radioactive and themselves decay into the nuclei of isotopes of other elements. This continues until the chain of successive nuclear transformations (*transmutations*) of one radioactive element into another is interrupted by the formation of a stable element. For instance, the final product of the decay of radioactive uranium and thorium is nonradioactive lead. The term for the spontaneous decay of the atomic nuclei of radioactive elements found in nature is *natural radioactivity*.

Since radioactive processes are unaffected by the environment, the age of uranium ore can be determined from the ratio of uranium to the final product, lead, contained in it. The age of uranium ore from various mines was found to be approximately the same: about  $4 \times 10^9$  years. This points to the conclusion that the crust of the Earth was formed about 4 billion years ago.

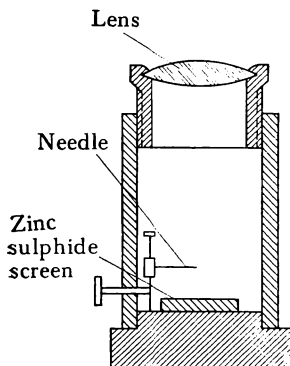
#### 40-4 Energy and Penetrating Power of Radioactive Radiation

One can obtain a notion of the energy of  $\alpha$ -particles emitted by the nuclei of radioactive elements by measuring their path in matter. The  $\alpha$ -particles ejected out of the nucleus at enormous initial velocities (up to 20 000 km/s) expend their energy on ionizing the atoms of the matter they penetrate and where they eventually are brought to a halt. In air  $\alpha$ -particles produce on the average 50 000 ion pairs per centimetre of their path.

The greater the energy of an  $\alpha$ -particle the greater the length of its path. The paths of  $\alpha$ -particles can be conve-

niently studied with the aid of a Wilson cloud chamber. The energy of  $\alpha$ -particles emitted in the course of natural radioactivity lies in the range from 4 to 9 MeV ( $1 \text{ MeV} = 10^6 \text{ eV}$ ). The path of an  $\alpha$ -particle in air may be from 2 to 12 cm, but only some micrometres in solids and liquids. Because of this  $\alpha$ -particles can be stopped by thin metal foil or even by a sheet of paper.

Fig. 40.6 Spinthariscopes.



The path of  $\alpha$ -particles can be determined with the aid of a *spinthariscopes* (Fig. 40.6). A spinthariscopes consists of a luminescent screen, a needle holding a radioactive sample and a lens. The latter is used to observe scintillations produced by  $\alpha$ -particles on the screen. The needle is moved away from the screen until there are no more scintillations to be seen. The corresponding distance of the needle from the screen can be assumed to be the maximum path of the  $\alpha$ -particles. Bringing the needle closer to the screen and counting the number of scintillations in each position one can plot the dependence of the number of  $\alpha$ -particles remaining in the flux on the distance, and from this find the average length of the path of the  $\alpha$ -particles. The use of a spinthariscopes made it possible to determine the rate of decay of radium nuclei: it was found to be  $3.7 \times 10^{10}$  atomic nuclei per gram per second.

The decay rate characterizes the *activity* of various radioactive samples. In the SI system the accepted unit of activity is the activity of a sample in which one atomic nucleus decays per second (1 decay/s). Units frequently used in practice are the curie and the rutherford.

One *curie* (c) is the activity of a sample in which there are  $3.7 \times 10^{10}$  atomic nuclei disintegrations per second:  $1 \text{ c} = 3.7 \times 10^{10} \text{ decay/s}$ . Hence, the activity of 1 g of radium is 1 c.

One *rutherford* (rd) is the activity corresponding to  $10^6$  decay/s:  $1 \text{ rd} = 10^6 \text{ decay/s}$ .

The velocities of electrons in  $\beta$ -radiation can be almost as high as the velocity of light, while their energies vary greatly: from about 0.01 to 2.3 MeV. The paths of electrons in substances are much longer than those of  $\alpha$ -particles, since the electrons produce much fewer ions in their wake and accordingly lose their energy much more slowly; in air at atmospheric pressure  $\beta$ -particles produce on the average about 50 ion pairs per 1-cm path. To screen  $\beta$ -radiation a metal layer of about 3-mm thick is required.

The energy of  $\gamma$ -quanta produced by radioactivity lies in the range from 0.02 to 2.6 MeV. The penetration power of  $\gamma$ -rays is much greater than that of X rays. To absorb the hardest  $\gamma$ -rays a layer of lead over 20-cm thick is required.

The intensity of  $\gamma$ -rays varies in inverse proportion to the square of the distance from the radiation source.

The intensity of irradiation by X and  $\gamma$ -rays is measured by the energy of radiation absorbed by a body. The unit for measuring *absorbed radiation* is the *roentgen* (R). One roentgen corresponds to a radiation energy the absorption of which by one kilogram of air under normal condition results in the production of ions with the total charge (irrespective of sign) of  $2.58 \times 10^4$  C. A brief irradiation of a human being with a dose of 20-50 R produces changes in his blood, a dose of 100-250 R causes radiation sickness, and a dose of 600 R is lethal.

### 40-5 Cherenkov Radiation

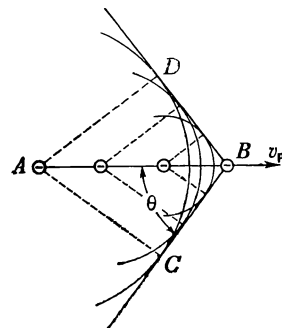
In 1934 the Soviet physicists Pavel A. Cherenkov (b. 1904) and Sergei I. Vavilov (1891-1951) discovered a new kind of luminosity subsequently called *Cherenkov radiation*. It can be observed when a source of radioactivity is surrounded by a dense transparent medium, for instance by water.

The radiation is produced when some particles, for instance electrons, move in a transparent medium at a speed exceeding the speed of light in this medium. This does not contradict the special theory of relativity, which says that the speed of a particle cannot exceed the speed of light in a vacuum,  $c = 3 \times 10^8$  m/s, since the speed of light in a medium is  $v_m = c/n$  (for water  $v = 2.25 \times 10^8$  m/s). Hence, an electron may move at a speed greater than  $v_m$  but less than  $c$ . In the case of water an electron would need an energy above 0.26 MeV.

Such motion of a particle is similar to the motion of a ship at a speed exceeding the velocity of wave propagation over the water: waves spreading in different directions are produced in the ship's wake. A similar phenomenon is observed in aircraft at supersonic speeds. The aircraft is followed by waves spreading with a cone-shaped front. The angle at the cone's apex decreases with the aircraft's speed.

A particle flying in a medium at a speed exceeding the speed of light (in that medium) excites its atoms, which emit coherent radiation travelling at a speed  $v_m$  not as high as the particle's speed  $v_p$ . A particle, having excited a wave at point A, covers the distance  $AB = v_p t$  in the time  $t$ , exciting waves at all points along its path (Fig. 40.7). The light wave excited at A propagates in a sphere to a distance  $AD = AC = v_m t$ . Similarly, waves emitted by the atoms

Fig. 40.7 Schematic representation of wavefront of Cherenkov radiation.



lying on the straight line  $AB$  propagate to shorter distances, forming a conical surface (with generatrices  $BC$  and  $BD$ ). Since  $\triangle ABC$  is a right triangle, it follows that

$$\sin \theta = \frac{v_m t}{v_p t} = \frac{v_m}{v_p}$$

It follows that the angle  $\theta$  decreases with the speed of the particle. Hence, Cherenkov radiation can be used to determine the speed of fast particles. This radiation can be observed in the water used in atomic reactors.

#### 40-6 Man-Made Transmutations

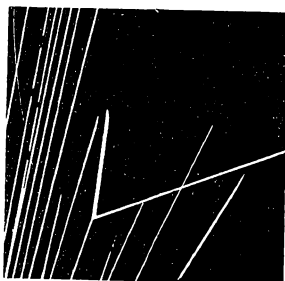
Studying the paths of  $\alpha$ -rays in various gases, in 1919 Rutherford made an important discovery. He observed scintillations in a spinthariscopes filled with air even when the distance between the source of  $\alpha$ -radiation and the screen exceeded the path of  $\alpha$ -particles in air. The scintillations ceased when oxygen or carbon dioxide were let in instead of air and started again when nitrogen was introduced.

Rutherford assumed that the scintillations were due to certain particles emitted by nitrogen nuclei hit by  $\alpha$ -particles. Subsequent research proved this to be true, and the particles emitted by the nitrogen nuclei were identified by their deflection in magnetic field as protons. It was established that in the case of a direct collision the  $\alpha$ -particle penetrates into the nitrogen nucleus; the nitrogen nucleus, having absorbed an  $\alpha$ -particle, loses its stability and, emitting a proton, turns into an oxygen nucleus. Thus, Rutherford observed the transmutation of helium and nitrogen nuclei into those of oxygen and hydrogen.

The transformation of the atomic nuclei of one element into those of another element was termed *nuclear reactions*. The great achievement of Rutherford was that he demonstrated the possibility of *artificial* nuclear reactions. Subsequently the British physicist Patric M. S. Blackett (b. 1897) took over 20 000 photographs of  $\alpha$ -particle tracks in a Wilson cloud chamber and of them eight were photographs of reactions described above. One such photograph is presented in Fig. 40.8: the track of one of the  $\alpha$ -particles terminates with a fork (the thick short track belongs to an oxygen nuclei and the thinner long one to a proton).

Rutherford and the British nuclear physicist Sir James Chadwick (1891-1974) also discovered other nuclear reactions initiated by  $\alpha$ -particles. In some of them the energy of the emitted protons turned out to be higher than that of ab-

Fig. 40.8 Transformation of nitrogen nucleus into oxygen nucleus: left track belongs to oxygen nucleus, right track to proton.



sorbed  $\alpha$ -particles. This points to the liberation of energy as the result of such a reaction. An example of a reaction of this sort is the transformation of an aluminium nucleus into a silicon nucleus, involving the absorption of an  $\alpha$ -particle and the emission of a high-energy proton.

The fact that the atomic nuclei emit protons in the above reactions and that a nuclear charge is always a multiple of the proton charge proves that protons belong to elementary particles making up the nuclei. However, if the nuclei were made up only of protons, their masses would be  $Z$  times that of the proton mass, where  $Z$  is the *atomic number*. Actually, nuclear masses are much greater. This means that atomic nuclei contain other particles besides protons.

### 40-7 The Neutron

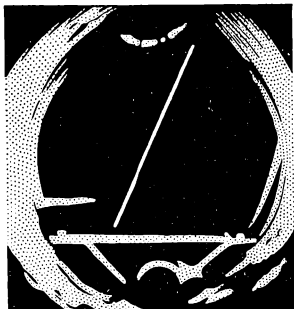
Experiments in which light elements were irradiated with  $\alpha$ -particles demonstrated that nuclear reactions were not always accompanied by the emission of protons. The German scientists Wather Bothe (1891-1957) and Howard S. Becker discovered in 1930 that the bombardment of beryllium by  $\alpha$ -particles produces a new type of radiation of very high penetration power. This was termed beryllium rays. This radiation left no tracks in a Wilson cloud chamber, did not produce scintillations and was not deflected in electric or magnetic fields, but it did knock hydrogen nuclei (protons) out of substances that contained hydrogen, and the nuclei of other elements, for instance of nitrogen out of its compounds. A similar radiation was also observed later when boron and some other elements were irradiated with  $\alpha$ -particles.

At first it was supposed that beryllium rays are  $\gamma$ -radiation. However, these rays penetrated through lead layers thick enough to have stopped all previously known  $\alpha$ -radiation. Moreover, calculations showed the energy of the photons corresponding to this radiation to be unrealistically high and, to make things worse, different depending on whether a proton, a nitrogen nucleus or some other nucleus had been knocked out. This cast doubt on the suggestion that beryllium radiation consisted of  $\gamma$ -rays.

In 1932 Chadwick suggested that beryllium rays consisted of neutral particles, with the mass of each close to that of the proton. He termed those particles *neutrons*. Subsequent research proved Chadwick's conjectures to have been true. This led to the discovery of another elementary particle, the neutron. Its rest mass is  $1.674\,92 \times 10^{-27}$  kg,



Fig. 40.9 Track of proton knocked out of paraffin by neutron.



a little more than that of the proton. Later collisions of neutrons with nuclei of various elements were registered on Wilson cloud chamber photographs. One such photograph is presented in Fig. 40.9. It depicts the track of a proton knocked out by a neutron (the neutron itself leaves no track).

Since the neutrons carry no charge, they do not interact with the electrons of the atoms and do not produce ions in their wake (a direct collision with an electron is an extremely rare event). This is the explanation for the high penetration power of a neutron flux. A neutron flies in a straight line until it hits a nucleus. It loses practically no energy in elastic collisions with heavy nuclei, bouncing off them like a ball bouncing off a wall. However, in collisions with light nuclei the neutron transmits to them an appreciable part of its energy, its speed decreasing after the collision. After several collisions its kinetic energy drops to one close to the energy of thermal motion of the particles of the medium. The term for such slow neutrons is *thermal neutrons*. The most efficient neutron moderators are substances containing hydrogen, for instance paraffin and water. Carbon is also a good moderator.

The probability of a neutron colliding with an atomic nucleus is much higher than the corresponding probability for an  $\alpha$ -particle because neutrons experience no electrostatic repulsion from the nuclei, as do  $\alpha$ -particles. In the course of inelastic collisions the neutrons easily penetrate into nuclei, initiating transmutations of numerous elements.

#### 40-8 Nuclear Structure. Nuclear Symbols and Reactions

The discovery of the neutron encouraged the German physicist Werner Karl Heisenberg (1901-1976) and the Soviet physicist Dmitrii D. Ivanenko (*b.* 1904) to put forward a hypothesis about the structure of the atomic nucleus according to which all atomic nuclei are made up only of protons and neutrons. These particles were given the common name *nucleons*.

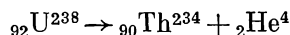
Since a nucleon's mass expressed in relative units is very close to unity (the proton mass is 1.007 276 and the neutron mass is 1.008 665), the mass of an atomic nucleus expressed in relative units is also very close to an integral number of nucleons it contains. This number is termed the *mass number* and denoted by the letter  $A$ . Since the number of protons in the nucleus is expressed by the atomic number  $Z$ , the number of neutrons is  $A - Z$ .

Almost the whole of the atom's mass is concentrated in its nucleus, which implies that its relative mass should be close to the integral number  $A$ . However, many elements exhibit a marked deviation from this rule. The causes of such deviations will be discussed in the following section.

In writing out nuclear reactions use is made of convenient notations showing the composition of the nucleus and its position in the Mendeleev Periodic Table. To denote the atomic nucleus use is made of the symbol of the respective chemical element. The atomic number  $Z$  is written to the left below the symbol and the mass number  $A$  to the right and above it. For instance  ${}_2\text{He}^4$  denotes a helium nucleus made up of four nucleons, two of which are protons and the other two neutrons.

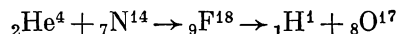
The symbol used for a free neutron outside the nucleus is  ${}_0\text{n}^1$  since its charge is zero, and for a proton the symbol used is  $\text{p}$  or  ${}_1\text{H}^1$ . The electron is denoted by symbols  $\beta^-$  or  ${}_{-1}\text{e}^0$ . The zero on top means that the electron's mass is small as compared with the mass of the nucleon and cannot affect the values of the mass numbers in nuclear reactions.

Using this notation one writes the reaction of radioactive decay of uranium involving the emission of  $\alpha$ -particles in the following form:



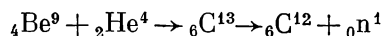
This reaction means that thorium and helium are produced as the result of the decay of uranium. It should be kept in mind that the number of nucleons and the charge are conserved in nuclear reactions, the only result of such reactions being the redistribution of nucleons among the nuclei. Accordingly, the sums of the upper indices in the left- and right-hand sides of the nuclear reaction should be equal. The same applies to the lower indices.

The reaction of proton creation in the course of the absorption of  $\alpha$ -particles by nitrogen nuclei (see Section 40-6) is written in the following form:

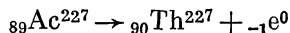


that is, the nitrogen nucleus absorbing an  $\alpha$ -particle turns into an unstable fluorine nucleus which, emitting a proton, turns into an oxygen nucleus.

The nuclear reaction equation for the case of the irradiation of beryllium by  $\alpha$ -particles (see Section 40-7) is written in the following form:



In the actinium-thorium transmutation  $\beta$ -particles are produced



But how can negatively charged electrons fly out of nuclei which contain only positive and neutral particles? Actually, the electron is created at the moment of nuclear decay as the result of the transformation of one of the nuclear neutrons into a proton:



(we shall see below that this equation does not give the complete picture of the neutron decay process). In the nuclei of nonradioactive elements the neutrons do not experience such a decay.

### 40-9 Isotopes

It was established as a result of studies of radioactive transmutations that there are atomic nuclei in nature with equal atomic numbers but with different mass numbers, for which the British scientist Frederick Soddy (1877-1956) suggested the term isotopes, since they occupy identical places in the Mendeleev Periodic Table. It was established that isotopes frequently appear in radioactive transmutations. For instance, radon nuclei can be of three types, with mass numbers 219, 220, 222; uranium, radium, thorium have several isotopes each, and so on.

However, it remained unclear whether other nonradioactive chemical elements also have isotopes. Is not the existence of isotopes an explanation for fractional relative atomic masses? For instance, the relative atomic mass of chlorine is 35.5; does not this mean that chlorine is a mixture of two or more isotopes?

The British scientist Sir Joseph John Thomson (1856-1940) was the first to start the hunt for isotopes of nonradioactive elements. Studying channel rays in a tube filled with neon, he discovered neon atoms with mass numbers 20 and 22. This was proof of the fact that nonradioactive elements, too, can have isotopes. Continuing Thomson's work the British physicist Francis W. Aston (1877-1945) in 1919 built an instrument capable of measuring atomic masses with an accuracy of up to 0.01 per cent. This made it possible to establish the existence of isotopes of numerous elements. The term for the instrument able of measuring atomic mass is *mass spectrometer*.

The design of a mass spectrometer is shown in Fig. 40.10. The instrument consists of a capacitor with plates  $P$  and magnets not shown in the figure. The upper diaphragms  $D_1$  and  $D_2$  admit into the capacitor a narrow beam of positive ions of the chemical element being studied.

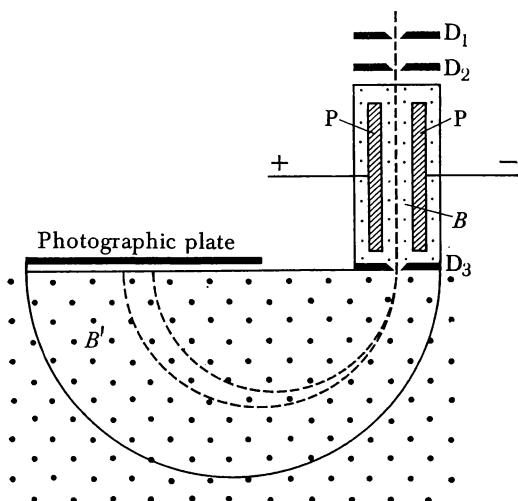


Fig. 40.10 Mass spectrometer.

Inside the capacitor there is an electric field which deflects the flying ions to the right and a magnetic field with an induction  $B$  (directed towards the reader) which deflects the ions to the left. In such fields only ions with a strictly defined velocity fly in a straight line, all others being deflected to the right or left. Thus, only ions with equal velocities (monochromatic ions) pass through the slit in the diaphragm  $D_3$ . Below the slit they enter a magnetic field with an induction  $B'$  (also directed towards the reader). This causes them to fly in a circle, the radius of the circle being the greater the greater the mass of the ions (see Section 25-18, formula (25.23)).

Completing a semicircle the ions strike a photographic plate. Depending on their mass, the ions of the isotopes move in different circles and strike different parts of the plate. The number of images of the slit  $D_3$  on the plate is equal to the number of isotopes. The positions of the images makes possible a very accurate estimate of the isotopic masses. The term for such photographs is *spectrograms* of isotopic masses.

Fig. 40.11 Mass spectrogram of germanium isotopes.

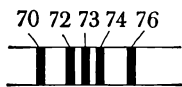


Figure 40.11 depicts a spectrogram of the isotopic masses of germanium. Germanium has five isotopes with mass numbers 70, 72, 73, 74 and 76. Obviously, a gram-atom of a mixture of these isotopes, depending on their concentrations, can be anything from 70 to 76. Actually, their concentrations in the Earth's crust are such that the relative atomic mass of germanium is 72.6.

Chlorine turned out to be a mixture of two isotopes, with mass numbers 35 (about 75 per cent) and 37 (25 per cent). Now we can understand why the relative atomic mass of chlorine is 35.5.

Mass-spectroscopic studies have shown that all chemical elements possess isotopes. Some of them are radioactive, others are stable. Several heavy elements only have radioactive isotopes with different half-lives.

As soon as it was established that atomic nuclei are made up of protons and neutrons, it became clear that the atomic nuclei of all isotopes of a chemical element have an equal number of protons but a different number of neutrons, and therefore have different masses. Since their atomic numbers  $Z$  are identical, the atoms of the isotopes have identical electron shells and have identical chemical and almost identical physical properties. Accordingly, isotopes cannot be separated by chemical methods; to separate them we rely on small differences in the rates of vapourization, diffusion, etc., resulting from the differences in their atomic masses.

Of great practical importance are the isotopes of hydrogen. In addition to the isotope  ${}_1\text{H}^1$  there exists also the so-called *heavy hydrogen*, or *deuterium*. Its atomic nucleus  ${}_1\text{H}^2$  is made up of one proton and one neutron; it is called a *deuteron* and is sometimes denoted  $D$ . The concentration of  ${}_1\text{H}^2$  is one for 6000 hydrogen atoms.

The term for the deuterium-oxygen compound  $\text{D}_2\text{O}$  is *heavy water*. Its density is  $1.108 \times 10^{-3} \text{ kg/m}^3$ ; it freezes at  $3.8^\circ\text{C}$  and boils at  $101.4^\circ\text{C}$ . A small number of  $\text{D}_2\text{O}$  molecules are always present in natural water. Heavy water can be separated with the aid of electrolysis. In the electrolysis of water the rate of decomposition of the  $\text{H}_2\text{O}$  molecules is higher because the  $\text{D}^+$  ions are heavier and less mobile than the  $\text{H}^+$  ions. Accordingly, water after electrolysis is enriched with heavy water.

There is also a third hydrogen isotope, *tritium*, denoted  $T$ . Its nucleus  ${}_1\text{H}^3$  is made up of one proton and two neutrons. It is radioactive with a half-life of 12.26 years.

The mass spectrometer was instrumental in the discovery of isotopes of the heaviest element known in the 1920s, uranium. Natural uranium is mainly a mixture of two isotopes:

${}_{92}\text{U}^{238}$  (with a half-life of  $4.5 \times 10^9$  years) and  ${}_{92}\text{U}^{235}$  ( $7 \times 10^8$  years), the concentration of  ${}_{92}\text{U}^{238}$  being 99.3 per cent and of  ${}_{92}\text{U}^{235}$  only about 0.7 per cent.

#### 40-10 Nuclear Forces

If atomic nuclei are made up solely of protons and neutrons, how does one explain their stability? Having like charges, protons placed very close together in the nucleus should repel each other with a terrific force. In spite of this, atomic nuclei are extremely strong particles.

For example, to split a helium nucleus into separate protons and neutrons an energy is required hundreds of thousands times greater than that required to separate both its electrons from the nucleus. This means that extremely powerful forces act between the nucleons inside the nucleus, exceeding many times over the electrical forces. These forces cannot be gravitational, acting in compliance with the law of universal gravitation, because gravitational forces are much weaker than the forces of electric repulsion of the protons. Consequently, nuclear forces are of a new type. They are the strongest of all the interactions known in nature.

Already Rutherford's experiments on the scattering of  $\alpha$ -particles by atomic nuclei described above produced the result that nuclear forces act over very short distances not exceeding  $10^{-14}\text{m}$ . The interaction of nucleons is studied in nucleon-nucleon scattering experiments. To study the forces of interaction of two protons or of a proton and a neutron, hydrogen nuclei are bombarded with protons and neutrons and the deflections of the scattered particles are investigated. Deuterons are also used as targets.

The experiments demonstrated that the forces of nuclear attraction act between any two nucleons at distances between their centres of about  $2 \times 10^{-15}\text{m}$ , falling off drastically with distance. At distances above  $3.0 \times 10^{-15}\text{m}$  they are already practically zero. On the other hand, when in the course of collisions the nucleons approach each other to within a distance of  $0.5 \times 10^{-15}\text{m}$ , the nuclear forces change their character to that of repulsion. Hence, the interaction of two nucleons is outwardly similar to the interaction between two molecules (see Sections 2-4 and 2-5; Figs. 2.2 and 2.3), but the forces and the energy of interaction of the nucleons are millions of times greater and the distances millions of times less.

Because of the very small radius of action of nuclear forces each nucleon in a nucleus comprising several nucleons

can interact only with its nearest neighbour but not with all the nucleons in the nucleus.

If this is the case the density of the matter in all nucleons should be approximately the same and should not rise with the increase in the number of nucleons making up the nucleus. Actually, the density of the nuclear matter both of light and heavy nuclei is almost the same, being about  $10^{17}$  kg/m<sup>3</sup>, that is, 1 cm<sup>3</sup> of nuclear matter weighs 100 million tons.

There is an apparent similarity between an atomic nucleus and a liquid drop. Nucleons in the nucleus, like molecules in a liquid, interact only with their nearest neighbours. The density of a nucleus, like that of a drop, is independent of its size. The surface nucleons are bonded unilaterally to the internal nucleons and surface tension should make the nucleus, just like a drop, assume a spherical form.

The nucleons in an excited nucleus vibrate like molecules in a heated drop. One of them may, as the result of numerous collisions, gain an energy high enough to overcome the attraction of nuclear forces and leave the nucleus, like a molecule evaporating out of a liquid. When a charged particle, for instance a proton or an  $\alpha$ -particle, is at a distance from the nucleus exceeding the radius of action of nuclear forces, the nucleus acts on the particle like a positively charged drop; in such conditions the nucleus does not act on a neutron.

The liquid drop model of the nucleus enables nuclear radii to be computed and provides a graphic explanation for some of its properties.

Experiments show the  ${}^4_2\text{He}$  helium nuclei to be especially strong. This is why  $\alpha$ -particles are often emitted as the result of the radioactive decay of the nuclei of heavy elements. Consequently, the strongest forces of attraction acting inside a nucleus are those between two protons and two neutrons. Generally, nuclei made up of equal numbers of protons and neutrons turn out to be the strongest, provided the number of protons is not too high. When the number of protons is high, the forces of electrical repulsion (which, in contrast to the nuclear forces, act between all protons in the nucleus and not only between neighbouring protons) decrease the strength of the nucleus, and so nuclei containing more neutrons than protons turn out to be more stable.

At present the nature of nuclear forces is not quite clear. It has been established that they are *exchange forces*. Exchange forces are of a quantum nature and have no analogy in classical physics. The nucleons are bonded by means of a third particle which they continuously swop. In 1935 the

Japanese physicist Hideki Yukawa (b. 1907) demonstrated that the theoretical values of the interaction forces coincide with experimental data if it is assumed that nucleons exchange particles with masses about 250 times the electron mass. These particles were subsequently termed  $\pi$ -mesons, or *pions*.

The predicted particles were actually discovered in 1947 by the British physicist Cecil Frank Powell (1903-1969) who studied cosmic rays at high altitudes using thick photoemulsion layer photographic plates.

The rest mass of the pion is about 270 times that of the electron. Pions are of three types: positive  $\pi^+$ , negative  $\pi^-$  and neutral  $\pi^0$ . Like nucleons interact via neutral  $\pi$ -mesons, while unlike nucleons interact via charged  $\pi$ -mesons. A proton and a neutron exchanging a charged  $\pi$ -meson experience continuous mutual transformations. A proton surrendering its positive  $\pi$ -meson itself becomes a neutron, while the initial neutron absorbs this  $\pi$ -meson and becomes a proton. The interaction involving a negative  $\pi$ -meson takes place along similar lines. Proof of the mutual transformations of neutrons and protons is obtained in experiments involving the scattering of a neutron flux by protons.

Free pions can be born as a result of a collision of a high-energy proton with another proton or neutron. They are produced during the bombardment of atomic nuclei with cosmic particles and protons in accelerators. However, over periods less than  $10^{-7}$  s the free pions disintegrate into other particles.

#### 40-11 Nuclear Binding

The nucleons in an atomic nucleus are held together by nuclear forces; therefore in order to split a nucleus into its constituent separate protons and neutrons a high energy must be expended. The term for this energy is the *binding energy* of the nucleus.

When free protons and neutrons associate to form a nucleus an equal energy is liberated. In compliance with Einstein's special theory of relativity, the mass of an atomic nucleus should be less than the mass of the free protons and neutrons from which it was formed. This mass difference,  $\Delta m$ , corresponding to the nuclear binding energy  $E_b$  is determined by the Einstein equation (see Section 39-9):

$$E_b = c^2 \Delta m \quad (40.1)$$



The nuclear binding energy is so great that it is quite possible to measure this mass difference directly. Such mass differences were actually observed for all atomic nuclei in mass spectrometers.

The term for the difference between the sum of the rest masses of the protons and neutrons making up the nucleus and the mass of the nucleus is the *mass defect* of the nucleus.

The binding energy is usually expressed in megaelectronvolts (1 MeV =  $10^6$  eV). Since an atomic mass unit (amu) is equal to  $1.66 \times 10^{-27}$  kg, one can find the energy corresponding to it:

$$E = mc^2, \quad E_{\text{amu}} = 1.66 \times 10^{-27} \times 9 \times 10^{16} \text{ J}$$

or

$$E_{\text{amu}} = \frac{1.66 \times 10^{-27} \times 9 \times 10^{16} \text{ J}}{1.6 \times 10^{-13} \text{ J/MeV}} = 931.4 \text{ MeV}$$

The binding energy can be measured directly from the energy balance in the reaction of nuclear spallation. The method based on the spallation of a deuteron by  $\gamma$ -rays was the first to be used for determining the deuteron's binding energy. However, the binding energy can be much more accurately determined from formula (40.1), because a mass spectrometer measures isotopic masses with an accuracy of  $10^{-4}$  per cent.

Let us by way of an example calculate the binding energy of the helium  ${}_2\text{He}^4$  nucleus ( $\alpha$ -particle). Its mass in atomic units is  $M({}_2\text{He}^4) = 4.001\,523$ ; the proton mass is  $m_p = 1.007\,276$ , the mass of a neutron is  $m_n = 1.008\,665$ . Hence, the mass defect of a helium nucleus is

$$\Delta m = 2m_p + 2m_n - M({}_2\text{He}^4)$$

$$\Delta m = 2 \times 1.007\,276 + 2 \times 1.008\,665 - 4.001\,523 = 0.030\,359$$

Multiplying by  $E_{\text{amu}} = 931.4 \text{ MeV}$ , we obtain

$$E_b = 0.030\,359 \times 931.4 \text{ MeV} \approx 28.3 \text{ MeV}$$

The masses of all isotopes have been measured in mass spectrometers and their mass defects and binding energies have been determined. The values of the binding energies of several isotopes are presented in Table 40.1. Such tables are used for calculating the energetics of nuclear reactions.

If the combined mass of the nuclei and particles produced as a result of some nuclear reaction is below the combined mass of the original nuclei and particles, this means that energy has been liberated in the course of the reaction corresponding to the decrease in the combined mass. When

Table 40.1 Binding energy of atomic nuclei

Nucleus	$E_b$ (MeV)	$E_b/A$ (MeV)	Nucleus	$E_b$ (MeV)	$E_b/A$ (MeV)
$^1_1\text{H}^2$	2.2	1.1	$^{26}_{26}\text{Fe}^{56}$	492.2	8.79
$^1_1\text{H}^3$	8.5	2.83	$^{30}_{30}\text{Zn}^{64}$	559.1	8.74
$^2_2\text{He}^3$	7.7	2.57	$^{50}_{50}\text{Sn}^{120}$	1020.6	8.50
$^2_2\text{He}^4$	28.3	7.075	$^{56}_{56}\text{Ba}^{138}$	1158.5	8.39
$^3_3\text{Li}^6$	32.0	5.33	$^{57}_{57}\text{La}^{139}$	1164.8	8.38
$^3_3\text{Li}^7$	39.2	5.60	$^{82}_{82}\text{Pb}^{206}$	1622.3	7.88
$^4_4\text{Be}^9$	58.2	6.47	$^{82}_{82}\text{Pb}^{208}$	1636.4	7.87
$^5_5\text{B}^{10}$	64.7	6.47	$^{86}_{86}\text{Rn}^{222}$	1708.2	7.69
$^5_5\text{B}^{11}$	76.2	6.93	$^{88}_{88}\text{Ra}^{226}$	1731.6	7.66
$^6_6\text{C}^{12}$	92.2	7.68	$^{89}_{89}\text{Ac}^{228}$	1741.6	7.64
$^6_6\text{C}^{13}$	97.1	7.47	$^{90}_{90}\text{Th}^{228}$	1743.0	7.64
$^7_7\text{N}^{14}$	104.7	7.48	$^{90}_{90}\text{Th}^{232}$	1766.5	7.61
$^8_8\text{O}^{16}$	127.6	7.975	$^{90}_{90}\text{Th}^{234}$	1777.7	7.60
$^8_8\text{O}^{17}$	131.8	7.75	$^{92}_{92}\text{U}^{233}$	1771.8	7.60
$^{10}_{10}\text{Ne}^{20}$	160.6	8.03	$^{92}_{92}\text{U}^{235}$	1783.8	7.59
$^{11}_{11}\text{Na}^{23}$	186.6	8.11	$^{92}_{92}\text{U}^{236}$	1790.2	7.586
$^{12}_{12}\text{Mg}^{24}$	198.3	8.26	$^{92}_{92}\text{U}^{238}$	1801.7	7.57
$^{13}_{13}\text{Al}^{27}$	225.0	8.33	$^{92}_{92}\text{U}^{239}$	1806.5	7.56
$^{14}_{14}\text{Si}^{30}$	255.2	8.51	$^{93}_{93}\text{Np}^{239}$	1807.0	7.56
$^{15}_{15}\text{P}^{30}$	250.6	8.35	$^{94}_{94}\text{Pu}^{239}$	1806.9	7.56
$^{15}_{15}\text{P}^{31}$	262.9	8.48	$^{94}_{94}\text{Pu}^{240}$	1813.3	7.555

the total number of protons and neutrons remains unaltered, the decrease in their combined mass is tantamount to an increase in the total mass defect as a result of the reaction and to the establishment of stronger bonds between the nucleons in the new nuclei as compared with the original nuclei. The energy liberated in the reaction is equal to the difference between the combined binding energy of the produced nuclei and the combined binding energy of the original nuclei. One can find this from the tables without calculating the total mass. This energy can be liberated into the surrounding medium in the form of the kinetic energy of the nuclei or the other particles and in the form of  $\gamma$ -quanta. Any spontaneous reaction serves as an example of an exoergic reaction.

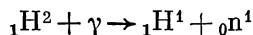
Let us calculate the energetics of the radium-radon transmutation reaction



The binding energy of the initial nucleus is 1731.6 MeV (see Table 40.1) and the combined energy of the nuclei formed is  $1708.2 + 28.3 = 1736.5$  MeV, exceeding the binding energy of the initial nucleus by 4.9 MeV. Consequently, the energy liberated in this reaction is 4.9 MeV and it is liberated mainly in the form of the kinetic energy of the  $\alpha$ -particle.

If nuclei and particles with a combined mass exceeding that of original nuclei and particles are formed in a reaction, such a reaction must necessarily involve the absorption of energy corresponding to the increase in the mass and can never proceed spontaneously. The amount of energy absorbed is equal to the difference between the combined binding energy of the original nuclei and the combined binding energy of the nuclei formed as a result of the reaction. This makes it possible for us to calculate the kinetic energy of a particle or nucleus bombarding the target nucleus required to realize this reaction or the magnitude of the  $\gamma$ -quantum required to split a nucleus.

Thus, the minimum energy of a  $\gamma$ -quantum capable of splitting a deuteron is equal to the deuteron's binding energy of 2.2 MeV, since a free proton and a free neutron ( $E_b = 0$ ) are formed in this reaction:



A good measure of agreement between theoretical calculations of this sort and experimental results proves the above explanation of the nuclear mass defect, and proves the principle of the proportionality of mass and energy established in the theory of relativity to be true.

It should be mentioned that reactions involving transformations of elementary particles (for instance,  $\beta$ -decay) are also accompanied by the absorption or liberation of energy corresponding to the variation of the combined mass of the particles.

An important characteristic of a nucleus is the average binding energy of the nucleus per nucleon,  $E_b/A$  (see Table 40.1). The greater it is the stronger are the bonds between the nuclei and the stronger is the nucleus. It can be seen from Table 40.1 that for most nuclei  $E_b/A$  is equal to about 8 MeV per nucleon, decreasing in the case of very light and very heavy nuclei. Among the light nuclei the helium nucleus occupies an exceptional position.

The dependence of  $E_b/A$  on the mass number of the nucleus,  $A$ , is shown in Fig. 40.12. In light nuclei most nucleons are on the surface of the nucleus, where they are unable to use all their bonds. Because of this  $E_b/A$  is not large. But

as the nuclear mass increases, the fraction of surface nucleons decreases with a resulting increase in  $E_b/A$ . However, the increase in the number of nucleons in the nucleus is accompanied by an increase in the Coulomb repulsion forces acting between the protons, thus weakening the bonds

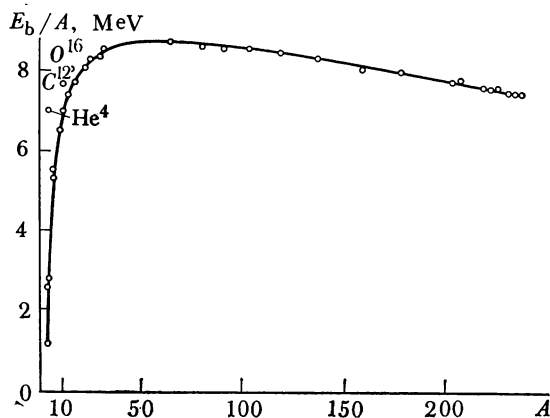


Fig. 40.12 Average binding energy of per nucleon versus mass number  $A$ .

the nucleus and reducing the magnitude of  $E_b/A$  in heavy nuclei. Hence,  $E_b/A$  is maximum for nuclei of medium mass (for  $A = 50-60$ ). These are the strongest nuclei.

This brings us to an important conclusion. The reactions of the fission of heavy nuclei into two medium nuclei, as well as the reactions of the synthesis of a medium or light nucleus from two lighter ones, produce nuclei with stronger bonds than the original (with greater  $E_b/A$ ). This means that energy is liberated in such reactions. This is the basic principle underlying the production of atomic energy, using the fission of heavy nuclei method (see Section 42-2) or of thermonuclear energy by the method of nuclear synthesis (see Section 42-6).

## Cosmic Rays. Elementary Particles

# 41

### 41-1 Cosmic Rays

In air free of ions a charged electroscope should retain its charge for an indefinite time. However, experiment shows it to lose its charge.

At first this phenomenon was attributed to the ionizing action of the radioactive radiation of the Earth. If this was true, then the ionizing radiation should become weaker with altitude. In 1912 it was established with the aid of balloons that the intensity of ionizing radiation in fact rises with altitude. This means that the source of the radiation is not on the Earth but somewhere in space. For this reason it was termed *cosmic rays*.

The studies of cosmic rays carried out in high altitude mountainous regions established that it consists of pions, protons, neutrons and other particles, among which many new previously unknown particles were discovered. These particles became known as *secondary particles*, because it was established that they were produced in the upper layers of the atmosphere as a result of interactions of *primary* cosmic particles coming out of space with the nuclei of atmospheric atoms.

It was established as a result of research that the intensity of cosmic rays near the poles of the Earth is about 1.5 times greater than at its equator. Primary cosmic radiation, on account of its deflection by the Earth's magnetic field, was identified as consisting of positively charged particles.

Much valuable information of primary cosmic radiation was obtained with the aid of artificial satellites and spacecraft.

At present primary cosmic radiation is known to consist of stable high-energy particles flying in space in all possible directions. The intensity of cosmic radiation in the solar system is on the average 2 to 4 particles per square centimetre per second. It consists mainly of protons ( $\sim 91$  per cent) and  $\alpha$ -particles (6.6 per cent) with small percentages of nuclei of other elements (under one per cent) and electrons ( $\sim 1.5$  per cent).

The average energy of cosmic particles is about  $10^4$  MeV, the energy of individual particles being as high as  $10^{12}$  MeV and more. It is, as yet, a mystery where the cosmic particles originate and where they are accelerated to such enormous energies. It is presumed that they are ejected in the outbursts of nova and supernova and are accelerated in the nonhomogeneous magnetic fields of interstellar space.

The Sun periodically (at the time of flares) emits solar cosmic rays consisting mainly of protons and  $\alpha$ -particles of low energy but high intensity. This fact has to be taken into account in drawing up plans for space flights.

The secondary particles also possess very high energies and, colliding with nuclei, breed new particles.

Figure 41.1 depicts a magnified picture of the disintegration of an atomic nucleus hit by a particle of high energy (about  $2 \times 10^8$  MeV). The track of the projectile is invisible (presumably it was a neutron). The nucleus split up into 17 particles, which have separated in various directions.

Avalanche-like breeding of particles in the upper layers of the atmosphere results in a cascade nuclear shower. Figure 41.2 shows an artificial cascade shower obtained in a Wilson cloud chamber with lead plate separators. A high-energy particle passing through the lead produces a shower of particles, which in passing through the next lead layer produce new showers.

The nuclear shower in the atmosphere fades out when the energy of the particle decreases to several tens of megaelectron-volts. The rest of their energy the protons spend on ionizing air; the neutrons are absorbed by the nuclei, inducing various nuclear reactions, and the pions, which constitute the bulk of the shower particles, decay. The photons and electrons produced in large numbers are strongly absorbed by the atmosphere.

A neutral pion very soon turns into two high-energy photons. The decay of charged pions produces new particles,  $\mu$ -mesons, or *muons*, which were discovered in 1938 in cosmic rays by the American physicist Carl D. Anderson (b. 1905), long before pions were discovered. The muon's mass is 207 times that of an electron, that is, it is  $3/4$  of the pion's mass. Muons are only of two kinds, positive and negative, denoted  $\mu^+$  and  $\mu^-$  respectively; while  $\pi^+$ -mesons decay into  $\mu^+$ -mesons,  $\pi^-$ -mesons decay into  $\mu^-$ -mesons.

Actually, in contrast to pions muons do not take part in nuclear interactions and spend their energy only on ionization. Because of this they have great penetration power and form the *hard component* of cosmic radiation. Muons penetrate the atmosphere and are observed even deep below the surface of the Earth.

Muons are unstable and exist only for some microseconds before decaying into other particles.

At sea level cosmic radiation is about 100 times less intense than at the fringe of the atmosphere and consists mainly of muons. The rest is electrons, photons and a small number of shower particles. From primary cosmic radiation only individual particles with exceptionally high energies (above  $10^7$  MeV) are able to pierce the atmosphere.

The muons and pions in cosmic rays fly at speeds close to the speed of light and because of relativistic time dilatation they are able to cover large distances before decaying (see Section 39-6).

Fig. 41.1 Disintegration of atomic nucleus hit by high-energy particle registered on nuclear emulsion plate.

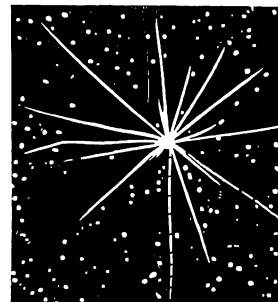


Fig. 41.2 Cascade shower caused by high-energy particle in Wilson cloud chamber.

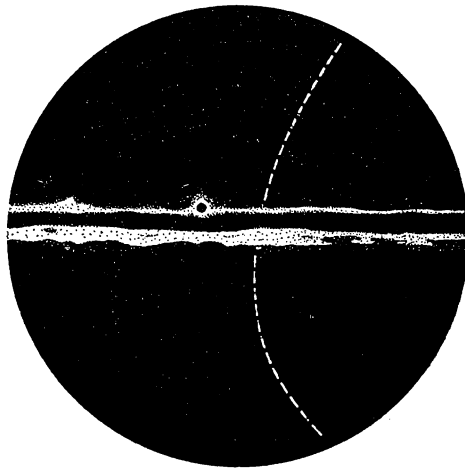


## 41-2 The Positron

In 1928 the British scientist Paul A.M. Dirac (b. 1902) predicted on the basis of his relativistic quantum theory the existence of a particle similar to the electron but carrying a positive charge. These particles later became known as *positrons*.

In 1932 Anderson discovered the tracks of positrons in cosmic radiation. He set up a strong magnetic field in a Wilson cloud chamber and observed tracks of small curvature which could be attributed to a high-energy positively

**Fig. 41.3** Positron track in Wilson cloud chamber (magnetic field pointing away from reader).



charged particle. To test his idea Anderson had to determine the precise direction of the particle's flight, for only by so doing could he use its deflection in the magnetic field to obtain its charge. He divided the cloud chamber by a lead plate. Passing through the plate the particle would lose its speed and move in a more curved path. On one of the photographs he again observed a track of this particle (Fig. 41.3). The direction of flight of the particle and its positive charge were now beyond doubt. (What would be the track left by an electron with the same energy?)

Calculations proved the mass and the magnitude of the charge of the new particle to be identical to those of the electron. Subsequent research demonstrated that the properties of the positron are identical to those of the electron with the exception of the charge sign. The notation for the positron is  ${}_{+1}e^0$ ,  $\beta^+$  or  $e^+$ .

The nuclear charge in  $\beta^+$ -decay decreases by unity and the element moves one place to the left in the Mendeleev Periodic Table.

### 41-3 The Neutrino

It was established in experiments that the  $\alpha$ -particles emitted by the nuclei of an isotope in the course of its  $\alpha$ -decay have a definite energy characteristic of the isotope which is easily calculated (see Section 40-11).

The situation is different in the case of  $\beta$ -decay. The  $\beta$ -particles emitted during the decay of identical nuclei have widely varying energies, all of them below some maximum level characteristic of the isotope. This maximum energy corresponds to the energy of electrons as calculated for a nuclear reaction from the law of energy conservation. So the question is: Where does the energy of the electrons flying at lower speeds go?

In 1930 the Austro-Swiss physicist Wolfgang Pauli (1900-1958) put forward a hypothesis according to which the nucleus in the act of  $\beta$ -decay emits two particles: an electron and a light neutral particle which takes away some of the energy liberated in the process. This particle leaves no trace, and for a long time efforts to detect it met with no success.

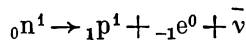
In 1933 the Italo-American physicist Enrico Fermi (1901-1954) developed a theory of  $\beta$ -decay and termed the particle the *neutrino*, which stands for "baby neutron". Still later nuclear reactions entailing the birth of positrons,  $\beta^+$ -decay, were discovered. In this case, too, a neutrino should be emitted by the nucleus simultaneously with a positron; this neutrino was denoted by  $\nu$ . The particle emitted by a nucleus together with an electron in the  $\beta^-$ -decay process was renamed *antineutrino* ( $\bar{\nu}$ ).

The neutrino and the antineutrino are similar, the only difference being that the spin of the antineutrino (the vector of its intrinsic angular momentum) coincides with the direction of its motion (i.e. it is a "right-hand screw" particle) and that of the neutrino is opposed to the direction of its motion (a "left-hand screw" particle). Both particles, like the photon, travel at the speed of light and have a zero rest mass.

The neutrino and antineutrino are born as a result of the decay of intranuclear nucleons. In the act of  $\beta^-$ -decay one of the neutrons inside the nucleus turns into a proton, an elec-

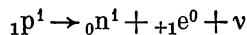


tron and an antineutrino:



The electron and the neutrino leave the nucleus, and the proton and the remaining nucleons form a new nucleus.

In the act of  $\beta^+$ -decay of a nucleus containing surplus protons one of them turns into a neutron, a positron and a neutrino being emitted simultaneously:



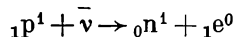
This reaction requires a supply of energy, since the mass of a proton is less than the mass of a neutron. For this reason the reaction can take place only inside a nucleus. A free proton is a stable particle.

In contrast to protons, free neutrons are liable to  $\beta^-$ -decay, since the neutron mass exceeds the combined mass of the proton and the electron. The half-life of free neutrons is about 12 minutes.

Not only the law of energy conservation but the law of the conservation of momentum as well is proof of the appearance of an antineutrino in the process of  $\beta^-$ -decay of free neutrons. Indeed, should a neutron disintegrate into only two particles—a proton and a neutron—the decay of a stationary (or a slow) neutron would result in the separation of both particles in opposite directions. Actually, the tracks of the proton and the electron in a Wilson cloud chamber form an angle. This means that a third particle is born in the process.

While the theory of  $\beta$ -decay was in good agreement with experimental results, all efforts to detect a neutrino or antineutrino were for a long time unsuccessful. The crux of the matter is that these tiny neutral particles do not practically interact with matter: flying past a nucleon or through it they are in contact with it for such a short time that in the majority of cases they are simply unable to interact with it. Because of this they have enormous penetration power and easily pierce the Earth and the Sun.

The antineutrino was found only in 1956. The American physicists C.L. Cowan, and F. Reines observed antineutrino capture by a proton:



resulting in the formation of a neutron and a positron.

A nuclear reactor served as a source of antineutrinos. Despite the fact that the probability of a proton capturing

an antineutrino is negligible, such events can still be observed because of the very large number of antineutrinos (of the order of  $10^{18}$  per second) born in the nuclear reactor.

Soon a neutrino was also observed.

#### 41-4 The Discovery of New Elementary Particles

In 1947 the English scientists George D. Rochester (b. 1908) and Clifford C. Butler (b. 1922), while studying cosmic rays, observed V-shaped tracks emerging from a point in a Wilson cloud chamber. It was obvious that they were the result of the decay of unknown particles which were neutral and left no tracks.

Subsequently these new particles were detected by other scientists as well. One of them has a mass one half of the proton's and is termed the *K-meson*, or *kaon*; the other is somewhat heavier than the proton and is termed  *$\Lambda$ -particle* (*lambda*).

In the following eight years they were joined by charged kaons as well as by two new kinds of heavy particles:  *$\Sigma$ -particles* (*sigma*) and  *$\Xi$ -particles* (*xi*).  *$\Sigma$ -* and  *$\Xi$ -particles*, like  *$\Lambda$ -particles*, turned out to be heavier than the proton and received the common name *hyperons*.

The discovery of kaons and hyperons was totally unexpected and for this reason they were termed *strange particles*. Their part in the structure of matter is still unclear, although it is obvious that all of them take part in nuclear interactions. Strange particles have several "mysterious" properties; for example, their lifetime is from the theoretical point of view unexpectedly high.

Elementary particles are born in collisions of high-energy particles with other particles. For a long time such collisions could be observed only in cosmic rays, which were the only source of high-energy particles. It was in cosmic rays that most elementary particles have been discovered.

At present accelerators for protons and other charged particles are being used to study elementary particles. The Serpukhov accelerator produces a beam of protons with an energy of  $76 \times 10^3$  MeV, as well as beams of other particles (pions, kaons, etc.) with energies of up to  $60 \times 10^3$  MeV. Gigantic accelerators designed to produce particles with energies of the order of  $10^6$  MeV are in the process of construction.

In the 1950s another kind of meson was discovered, the  $\eta$ -meson (eta), as well as the heaviest particle, the  $\Omega^-$ -hyperon (omega).

In 1961-2 experimental proof was produced of the existence of a second kind of neutrino—the *muonic neutrino*, denoted  $\nu_\mu$ . The notation for the electronic neutrino was accordingly changed to  $\nu_e$ .

Muons are born together with their neutrinos in the process of the decay of charged pions

$$\pi^+ \rightarrow \mu^+ + \nu_\mu$$

$$\pi^- \rightarrow \mu^- + \bar{\nu}_\mu$$

The properties of the muon neutrino ( $\nu_\mu$ ) and anti-neutrino ( $\bar{\nu}_\mu$ ) are quite similar to those of the electronic neutrino ( $\nu_e$ ) and antineutrino ( $\bar{\nu}_e$ ), but experiment has proved them to be different particles.

A wonderful property of the muon, as yet unexplained, is its absolute similarity with the electron in everything but its mass; the muon is 207 times heavier than the electron. This “heavy electron” is even able to temporarily occupy the place of the electron in an atom, with an orbit very close to the nucleus. The name given to such an atom is *mesoatom*.

The decay of muons results in the formation of electrons and positrons and of two neutrinos, electronic and muonic:

$$\mu^- \rightarrow e^- + \bar{\nu}_e + \nu_\mu$$

$$\mu^+ \rightarrow e^+ + \nu_e + \bar{\nu}_\mu$$

#### 41-5 Classification of the Elementary Particles

The most important property common to all elementary particles is their capacity for mutual transformation. During the decay of particles some vanish, others are born. Mutual transformation also takes place in the collision of two high-energy particles. For instance, two protons may upon collision turn into other particles

$$p + p \rightarrow p + n + \pi^+$$

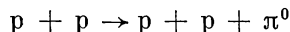
$$p + p \rightarrow p + \Lambda + K^+$$

All the transformations comply to the total energy conservation law, which includes the energy corresponding to the rest masses and the kinetic energy of the particles. These energies are mutually convertible.

The spontaneous decay of a particle results in the formation of particles whose combined rest mass is less than that of the particle which experienced decay, the energy corresponding to this difference in rest masses being transformed into the kinetic energy of the particles produced in the decay act.

A collision between two particles can result in a reverse energy transformation. In the examples cited above the rest mass of the newly born particles exceeds the rest mass of two colliding protons, the difference being gained at the expense of their kinetic energy.

Two colliding protons may give birth to a pion



if their kinetic energy prior to the collision exceeds the energy corresponding to the pion's rest mass. This example is especially impressive because the original particles are retained and a new one is born as well.

Transformations of particles comply, in addition, to the law of total energy conservation, as well as the laws of the conservation of charge and momentum. The elementary particles to this day are tabulated in Table 41.1.

The most important characteristic of a particle is its mass. This reflects its inertial and gravitational properties and determines its energy reserve. Table 41.1 shows the values of rest masses expressed in megaelectron-volts. The lightest particle with a rest mass is the electron (0.511 MeV).

Most particles have a spin (intrinsic angular momentum). We can see them as spinning about their axes like tops. Particles of each kind have a quite definite spin: assuming the spin of a photon to be unity, the spins of all particles are 0, or 1/2 or 1 (except the  $\Omega^-$ -hyperon whose spin is 3/2).

Some particles are neutral, others carry positive or negative electric charges equal to the electron charge in magnitude. The electric charge is included in the notation of all particles except the proton.

Almost all elementary particles are unstable. Only the proton, the electron and particles having no rest mass (photon and neutrino) are stable in the free state. Other particles decay spontaneously, and all of them except the neutron have very short mean lifetimes. Table 41.1 shows typical decay modes.

The elementary particles are divided into four classes:

(1) photons ( $\gamma$ -quanta), which have no rest mass and no electric charge and whose spin is 1;

**Table 41.1** Elementary particles

Category	Name	Symbol		Mass (MeV)	Spin	Charge		Mean life (s)	Principal decay modes
		particle	anti- particle			particle	anti- particle		
Photon	Photon, $\gamma$ -quantum	$\gamma$	$\gamma$	0	1	0	0	stable	
Leptons	Electron, positron	$e^-$	$e^+$	0.511	1/2	-1	+1	stable	
	Electronic neutrino	$\nu_e$	$\bar{\nu}_e$	0	1/2	0	0	stable	
	Muon	$\mu^-$	$\mu^+$	106	1/2	-1	+1	$2.2 \times 10^{-6}$	$\mu^- \rightarrow e^- + \bar{\nu}_e + \nu_\mu$
	Muonic neutrino	$\nu_\mu$	$\bar{\nu}_\mu$	0	1/2	0	0	stable	
Mesons	Pions	Neutral pion	$\pi^0$	135	0	0	0	$0.8 \times 10^{-16}$	$\pi^0 \rightarrow 2\gamma$
		Charged pion	$\pi^+$	140	0	+1	-1	$2.6 \times 10^{-8}$	$\pi^+ \rightarrow \mu^+ + \nu_\mu$
	Kaons	Neutral kaon	$K^0$	498	0	0	0	$0.9 \times 10^{-10}$	$K^0 \rightarrow \pi^+ + \pi^-$ $K^0 \rightarrow 2\pi^0$
		Charged kaon	$K^+$	494	0	+1	-1	$1.2 \times 10^{-8}$	$K^+ \rightarrow \mu^+ + \nu_\mu$ $K^+ \rightarrow \pi^+ + \pi^0$ $K^+ \rightarrow \pi^+ + \pi^- + \pi^+$
	Eta-meson	$\eta$	$\eta$	549	0	0	0	$2.4 \times 10^{-19}$	$\eta \rightarrow 2\gamma$ $\eta \rightarrow 3\pi^0$ $\eta \rightarrow \pi^+ + \pi^- + \pi^0$

Continuation									
Category	Name	Symbol		Mass (MeV)	Spin	Charge		Mean life (s)	Principal decay modes
		particle	anti- particle			particle	anti- particle		
Nucleons	Proton	p	$\bar{p}$	938.2	1/2	+1	-1	stable	$n \rightarrow p + e^- + \bar{\nu}_e$
	Neutron	n	$\bar{n}$	939.6	1/2	0	0	$0.93 \times 10^3$	
Barions	Lambda-hyperon	$\Lambda$	$\bar{\Lambda}$	1116	1/2	0	0	$2.5 \times 10^{-10}$	$\Lambda \rightarrow \begin{cases} p + \pi^- \\ n + \pi^0 \end{cases}$
		$\Sigma^+$	$\bar{\Sigma}^+$	1189	1/2	+1	-1	$0.8 \times 10^{-10}$	$\Sigma^+ \rightarrow \begin{cases} p + \pi^0 \\ n + \pi^+ \end{cases}$
	Sigma-hyperon	$\Sigma^0$	$\bar{\Sigma}^0$	1192	1/2	0	0	$10^{-14}$	$\Sigma^0 \rightarrow \Lambda + \gamma$
		$\Sigma^-$	$\bar{\Sigma}^-$	1197	1/2	-1	+1	$1.5 \times 10^{-10}$	$\Sigma^- \rightarrow n + \pi^-$
	Xi-hyperon	$\Xi^0$	$\bar{\Xi}^0$	1315	1/2	0	0	$3 \times 10^{-10}$	$\Xi^0 \rightarrow \Lambda + \pi^0$
		$\Xi^-$	$\bar{\Xi}^-$	1321	1/2	-1	+1	$1.7 \times 10^{-10}$	$\Xi^- \rightarrow \Lambda + \pi^-$
Hyperons	Omega-hyperon	$\Omega^-$	$\bar{\Omega}^-$	1672	3/2	-1	+1	$1.3 \times 10^{-10}$	$\Omega^- \rightarrow \begin{cases} \Lambda + K^- \\ \Xi^0 + \pi^- \\ \Xi^- + \pi^0 \end{cases}$

- (2) leptons—light particles whose spins are  $1/2$ ;
- (3) mesons—intermediate particles whose spins are 0;
- (4) barions—heavy particles (the lightest barion is the proton) whose spins, except that of the  $\Omega^-$ -hyperon, are  $1/2$ .

The particles of various categories are distinguished not only by their mass and spin. For instance, photons and leptons take no part in nuclear interactions, while mesons and barions do.

With leptons and barions the law of the conservation of the number of particles is satisfied. Thus, when one barion vanishes another is born in its place. The law of particle number conservation is responsible for the stability of the proton: it is the lightest barion and on the strength of this cannot decay spontaneously and give birth to another barion. The barion and the lepton number conservation laws have been tested in experiment on numerous occasions.

Photons and mesons do not obey the particle number conservation law and they may appear and vanish in any number.

#### **41-6 Antiparticles. Mutual Transformation of Substance and Field**

It follows from relativistic quantum theory that for each particle there should be a corresponding antiparticle, that is, a similar particle with the same mass, spin, lifetime but with a charge of opposite sign, with different mutual orientation of magnetic moment and spin, and with several other different characteristics.

The first antiparticle to be discovered was the “positive electron”, the positron. Negative and positive muons and negative and positive pions and kaons are also examples of such particle-antiparticle pairs. The name for an antiparticle is obtained by adding the prefix anti to the name of the respective particle. The same notation is used for the antiparticle, but with a bar on top (see Table 41.1). The photon, the neutral pion and the eta-meson have no antiparticles (in this case the particle may be said to coincide with the antiparticle).

Like their counterparts the antiproton, positron and antineutrino are stable, the other antiparticles being unstable. Table 41.1 describes the decay processes of such particles. Antiparticles disintegrate into corresponding antiparticles.

The absorption of  $\gamma$ -quanta with energies above 1 MeV has been found to be accompanied by the generation of

electron-positron pairs. When a  $\gamma$ -quantum flies through the strong electric field in the vicinity of the nucleus, it is transformed into an electron-positron pair:

$$\gamma \rightarrow e^- + e^+$$

The generation of electron-positron pairs can be seen when  $\gamma$ -radiation passes through a lead plate dividing a Wilson cloud chamber. The tracks of the positrons and the electrons diverge in symmetrical curves in the form of a V (Fig. 41.4).

Since the energy corresponding to the rest mass of the electron or the positron is 0.511 MeV,  $\gamma$ -quanta can be transformed into electron-positron pairs only if their energy exceeds 1.02 MeV.

Very high energy  $\gamma$ -quanta, produced as a result of the decay of neutral pions forming part of secondary cosmic radiation, generate electrons and positrons which also have high energies and in interaction with the atmosphere emit  $\gamma$ -bremsstrahlung. This, in turn, generates new pairs, and so on. This is the process that is responsible for the formation of the so-called *soft component of secondary cosmic radiation*, intensively absorbed by the atmosphere.

If electrons and positrons can be born out of  $\gamma$ -quanta, they can, obviously, also disintegrate into  $\gamma$ -quanta. Experiments by the French nuclear physicists Irène (1897-1956) and Frédéric (1900-1958) Joliot-Curie demonstrated the annihilation of the positron and the electron, accompanied in most cases by the generation of two  $\gamma$ -quanta with energies of 0.51 MeV separating in opposite directions (sometimes three  $\gamma$ -quanta with a total energy of 1.02 MeV are formed):

$$e^- + e^+ \rightarrow 2\gamma$$

Other examples of such transformations can also be cited. The decay of a neutral pion is accompanied by the formation of two  $\gamma$ -quanta:

$$\pi^0 \rightarrow \gamma + \gamma$$

that is, the energy corresponding to the rest mass of the pion is transformed into electromagnetic radiation energy.

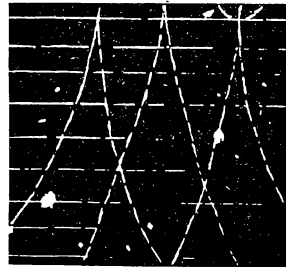
The collision of a high-energy  $\gamma$ -quantum with a proton can produce a neutron and a pion:

$$\gamma + p \rightarrow n + \pi^+$$

In this case the rest mass increases at the expense of the energy of electromagnetic radiation.

These experiments prove that electromagnetic radiation, whose quanta (photons) have no rest mass, can be transformed into particles with rest mass and vice versa.

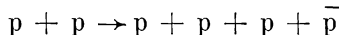
Fig. 41.4 Tracks of electron-positron pairs; three  $\gamma$ -quanta transformed into pairs inside lead (magnetic field pointing away from reader).



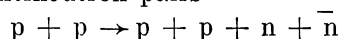


All the facts discussed above confirm the point that matter exists in the form of bulk substance and field and that both forms of matter are mutually convertible. Such transformations may involve kinetic energy. For instance, a proton may gain kinetic energy in the electric field of an accelerator and, colliding with another proton, produce new particles at the expense of its kinetic energy.

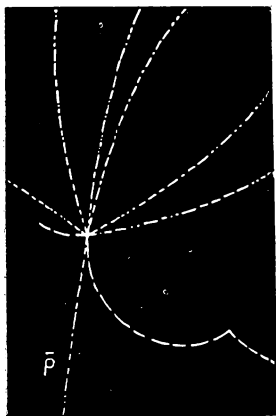
In 1955 the American physicist Ernest O. Lawrence (1901-1958) and his collaborators produced the antiproton and in 1956 the antineutron. The particles were produced in a high-energy accelerator in which protons with an energy of  $6 \times 10^3$  MeV bombarded target protons. Colliding protons gave birth to proton-antiproton pairs



and to neutron-antineutron pairs



**Fig. 41.5** Proton-antiproton annihilation.



An antiproton meeting a proton and an antineutron meeting a neutron annihilate each other, giving birth to several neutral and charged pions (the average number being five). Figure 41.5 illustrates the annihilation of an antiproton-proton pair in a bubble chamber. The antiproton  $\bar{p}$  moving from below meets a proton. In this case four positive and four negative pions which separated in different directions (the magnetic field is directed away from the reader) were produced in the act of annihilation. The kink in one track in the lower part of the figure is due to the pion's decay:  $\pi^+ \rightarrow \mu^+ + \nu_\mu$  (the neutrino leaves no track).

Neutral pions decay into  $\gamma$ -quanta. Charged pions decay giving birth to muons and neutrinos; muons, in turn, decay giving birth to electrons, positrons and neutrinos. The final act in the nucleon-antinucleon annihilation process is the annihilation of positron-electron pairs, the final products being several  $\gamma$ -quanta and neutrinos.

The discovery of antinucleons pointed to the possibility of the existence of *antimatter*, made up entirely of antiparticles. Thus, a negatively charged antiproton with a positron in its orbit is antihydrogen. Antinucleons can also form the nuclei of other antiatoms (up to now only antideuteron and the nucleus of antihelium have been obtained). Clearly, the production of antimatter is fraught with colossal difficulties, since it annihilates in contact with matter. There is a possibility of whole antiworlds made of antistubstance existing somewhere in the universe. However, this has not, as yet, been established as fact.

### 41-7 The Quark Model

Table 41.1 includes 35 elementary particles and antiparticles. All these particles are part of the structure of substance (or of antistubstance). They determine the forces of interaction between particles of other sorts and take part in the acts of transformation of particles.

In the sixties a big "family" of very short-lived particles was discovered. They were named *resonance particles*, or *resonances*. The lifetime of resonances is so small ( $10^{-22}$ – $10^{-23}$  s) that they cannot be regarded as real particles. They decay into other particles, leaving no traces, and can be recorded only indirectly.

Up to now the number of known elementary particles has reached about 200 (including resonances, the number of which is rising continuously). Clearly, the term elementary particle has lost its former meaning. A particle is classified as elementary if there is no proof that it is made up of other particles. The possibility should not be overlooked that many elementary particles may turn out to be made up of more elementary particles (like the atoms which until the beginning of the twentieth century were considered to be elementary particles).

In 1964 the American physicist Murray Gell-Mann (b. 1929) advanced a hypothesis according to which all mesons and barions are made up of fundamental particles of three kinds, which he called *quarks*, and of their antiparticles. They were termed *p-quark*, *n-quark* and  *$\lambda$ -quark*; antiquarks were denoted by a bar on top:  $\bar{p}$ ,  $\bar{n}$  and  $\bar{\lambda}$  respectively. All quarks have the same spin, equal to  $1/2$ . An unusual property of quarks is that they have a fractional electric charge (the charge of the positron being unity).

p-quark: $+2/3$	$\bar{p}$ -quark: $-2/3$
n-quark: $-1/3$	$\bar{n}$ -quark: $+1/3$
$\lambda$ -quark: $-1/3$	$\bar{\lambda}$ -quark: $+1/3$

Three quarks in different combinations are supposed to make up any barion, with a combined charge of 0,  $+1$ , or  $-1$ . The spins of three quarks constituting the  $\Omega$ -hyperon point in one direction, giving a combined spin of  $3/2$ . In other barions one of the spins is directed against two others, giving combined spins of  $1/2$ . Antiquarks are not present in barions—they make up antibarions.

All mesons are made up of one quark and one antiquark. For instance, the combination  $\bar{n}p$  produces a positive pion

and  $n\bar{p}$  a negative pion;  $\bar{\lambda}p$  produces a positive kaon and  $\lambda\bar{p}$  a negative kaon. The spins of quarks and antiquarks making up a meson are opposite and its spin is zero.

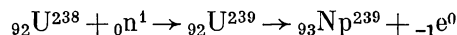
The mass of quarks was calculated to be equal to  $(5-10) \times 10^3$  MeV, that is, from 5 to 10 nucleon masses. Hence, the binding energy of quarks inside a nucleon should be extraordinarily high, since the mass defect turns out to be over 90 per cent of the rest mass of free quarks.

The quark model of the structure of elementary particles agrees well with experiment; however, no one has, as yet, been able to detect a quark.

## 42 Nuclear Power and Its Utilization

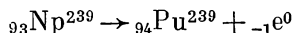
### 42-1 Transuranium Elements

The first element with an atomic number above 92 was produced in 1940 by the American scientists E.M. McMillan and P.H. Abelson when they bombarded uranium with neutrons. The isotope  ${}_{92}\text{U}^{238}$ , after absorbing a neutron, is transformed into the isotope  ${}_{92}\text{U}^{239}$ , which after  $\beta$ -decay turns into a new chemical element termed *neptunium* (Np):



These experiments demonstrated the possibility of the existence of elements heavier than uranium termed *transuranium elements*.

The newly discovered isotope  ${}_{93}\text{Np}^{239}$  exhibited  $\beta$ -radioactivity with a half-life of 2.3 days. It decays into the next transuranium element, plutonium:



The  ${}_{94}\text{Pu}^{239}$  isotope exhibits  $\alpha$ -radioactivity, but its half-life is  $2.44 \times 10^4$  years. This makes its accumulation in great quantities possible, a fact of major importance for the production of nuclear energy.

The following years witnessed the discovery of isotopes of transuranium elements of ever increasing atomic number, produced by irradiation of heavy nuclei with neutrons,  $\alpha$ -particles and heavy ions. There are great technical problems involved in the production of transuranium elements, the major one being that the half-lives of the isotopes fall off drastically with increasing  $Z$ .

Claims to the discovery of elements with  $Z = 104$  and  $Z = 105$  have been filed by an American and by a Soviet group, and the elements have not yet been officially named. The Soviet group proposed the name *kurchatovium* for the element 104 ( ${}_{104}\text{Ku}^{260}$ ) after Igor V. Kurchatov (1903-1960), the Soviet physicist who studied neutron reactions, whereas the American group gave the name *rutherfordium* ( ${}_{104}\text{Rf}^{257}$ ) to another isotope of the same element. For the element 105 the Soviet and American names are *bohrium* (Bh) and *hahnium* ( ${}_{105}\text{Ha}^{260}$ ) respectively.

## 42-2 Fission

In the 1930s laboratories in many countries were engaged in experiments involving the irradiation of natural uranium with neutrons. In 1938 the German scientists Otto Hahn (1879-1968) and Fritz Strassmann (b. 1902), analyzing chemically pure uranium irradiated with neutrons, discovered barium and lanthanum. The appearance of these elements, which occupy places in the middle of the Mendeleev Periodic Table, was a puzzling fact.

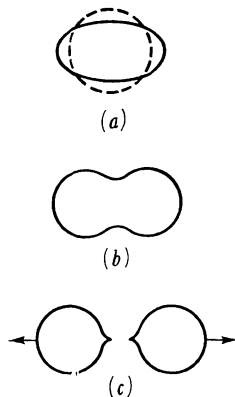
The Austrian physicists Lise Meitner (1878-1968) and Otto R. Frisch (1904-1979) explained it as disintegration of uranium nuclei into two approximately equal parts. This phenomenon was termed *fission* and the newly formed nuclei were termed *fission fragments*.

It was mentioned above (see Section 40-11) that energy should be liberated in the process of the fission of heavy nuclei. The average binding energy per nucleon,  $E_b/A$ , because of Coulomb repulsion forces acting between protons, is almost 1 MeV lower in heavy nuclei than in nuclei of medium mass (see Fig. 40.12). Since over 200 nucleons are involved in one act of fission, the total energy liberated in the act of fission of one heavy nucleus is about 200 MeV. This agrees with experimental data.

Bohr attributed the nuclear fission of natural uranium to its isotope  ${}_{92}\text{U}^{235}$  and this was confirmed in 1940. Nuclei of  ${}_{92}\text{U}^{235}$  which have absorbed neutrons turn into nuclei of  ${}_{92}\text{U}^{236}$ , which very quickly disintegrate into two approximately equal parts.

A graphic physical picture of fission is provided by a model representing the nucleus in the form of a positively charged liquid droplet (the *liquid drop model* of a nucleus). The nucleus, having absorbed a neutron, is in an excited state, since when a neutron is captured by a nucleus its binding energy is liberated (7.6 MeV for  ${}_{92}\text{U}^{236}$ ); when

**Fig. 42.1** Schematic representation of fission of heavy nuclei. Excited nucleus vibrates changing its shape: (a) little deformed nucleus; (b) greatly deformed nucleus; (c) nuclear fission.



a fast neutron is absorbed the nucleus gains, in addition, its kinetic energy. The excited nucleus starts vibrating like a drop of mercury changing its shape. When the excitation energy is not too high, the forces of surface tension are able to restore the nucleus to its spherical shape (Fig. 42.1a). If, on the other hand, the nucleus is greatly excited, its deformation may be so great (Fig. 42.1b) that at some moment the Coulomb repulsion forces acting between both parts of the nucleus will prevail over the nuclear binding forces and the nucleus will split into two parts flying in opposite directions (Fig. 42.1c). The fragments are rarely equal, one of them usually being about one and a half times larger than the other.

It was established that the nuclei of  ${}_{92}\text{U}^{238}$  are also liable to fission, but fast neutrons with energies above 1.1 MeV are required for this. Otherwise the excitation energy of the newly formed  ${}_{92}\text{U}^{239}$  will not be high enough for fission and instead nuclear reactions described in the preceding section will take place.

### 42-3 Chain Reactions

Since the neutron concentration in the nuclei of heavy atoms is much greater than in the nuclei of atoms occupying the middle of the Mendeleev Periodic Table, the fission fragments are greatly overloaded with neutrons. Accordingly, neutrons are liberated in the process of fission of heavy nuclei. It was demonstrated in experiments that two or three neutrons are liberated in the act of fission of one  ${}_{92}\text{U}^{235}$  nucleus (on average, 2.5 neutrons per act of fission). These secondary neutrons may spark off the fission of other nuclei and a *fission chain reaction*, proceeding in the absence of neutron irradiation of uranium, may be the result (Fig. 42.2).

However, in actual conditions by no means all the neutrons born in the process of fission take part in the fission of other nuclei. Some of them are captured by the nuclei of foreign atoms incapable of fission, others leave the uranium block (neutron leakage).

The term for the ratio of the number of fission acts produced by secondary neutrons to the number of fission acts in which they themselves were born is *effective neutron breeding coefficient* ( $K_{\text{eff}}$ ). For  ${}_{92}\text{U}^{235}$  it would be equal to 2.5 if all the secondary neutrons take part in the fission of new nuclei.

For  $K_{\text{eff}} < 1$  each new generation of neutrons sparks off an ever decreasing number of fission acts and in the absence of an external source of neutrons the reaction soon dies down. At  $K_{\text{eff}} = 1$  the rate of fission is held at a constant level. Such a course of self-sustaining chain reaction is termed *critical* and is maintained in nuclear reactors. At  $K_{\text{eff}} > 1$  each new generation of neutrons produces an

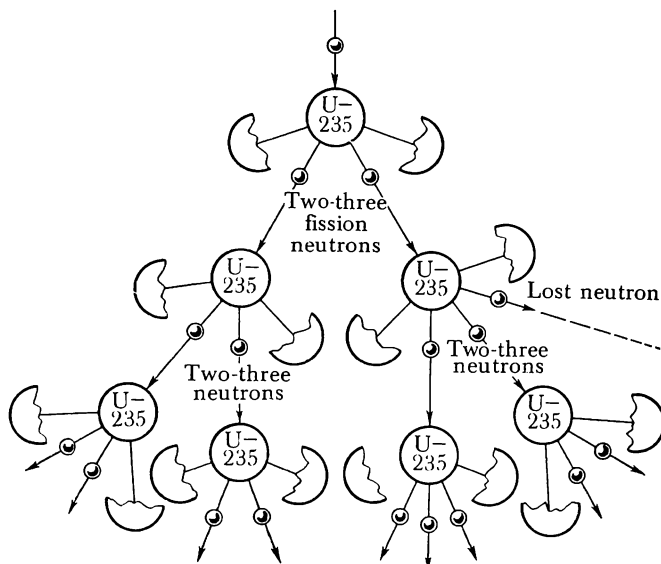


Fig. 42.2 Schematic representation of chain reaction of fission of  ${}_{92}\text{U}^{235}$ .

increasing number of fission acts and the chain reaction assumes the character of an avalanche. Since the neutrons liberated in the process of fission are instantly (in a time of  $10^{-7}$ - $10^{-8}$  s) captured by other uranium nuclei, causing their fission, such a chain reaction builds up rapidly, developing into a blast accompanied by the liberation of enormous energy and a rise in the temperature of the surrounding medium of several million degrees. Such a chain reaction takes place during the explosion of an atom bomb.

It can be easily calculated that, if the fission of one nucleus liberates an energy of 200 MeV, the fission of  $2.6 \times 10^{21}$  nuclei contained in one gram of uranium will liberate an energy of about  $8.3 \times 10^{10}$  J—an equivalent of the energy liberated in the process of the combustion of three tonnes of coal.

Nuclear chain reactions can be realized with the isotopes  ${}_{92}\text{U}^{235}$ ,  ${}_{92}\text{U}^{233}$  and  ${}_{94}\text{Pu}^{239}$ . These materials are termed

*nuclear fuel*, or *fissile materials*. Neutrons of any energy including the slow (thermal) neutrons are able to cause the fission of such materials.

Only one of the fissile materials,  ${}_{92}\text{U}^{235}$ , exists in nature. Its concentration in natural uranium is 0.7 per cent. The main isotope of natural uranium,  ${}_{92}\text{U}^{238}$ , cannot set off a chain reaction. Two to three neutrons (2.3 on average) are born in the act of fission of a  ${}_{92}\text{U}^{238}$  nucleus, the same as in the case of  ${}_{92}\text{U}^{235}$ , but of them, on average, one has insufficient energy for fission and of the rest only 1/5 do not lose their energy in collisions and are able to initiate the fission of new  ${}_{92}\text{U}^{238}$  nuclei. It can be easily calculated that the neutron breeding coefficient for  ${}_{92}\text{U}^{238}$  cannot exceed 0.3 and a chain reaction is impossible.

The other two fissile materials are synthesized:  ${}_{94}\text{Pu}^{239}$  from  ${}_{92}\text{U}^{238}$  after successive transformations, described in Section 42-1, and  ${}_{92}\text{U}^{233}$  from  ${}_{90}\text{Th}^{232}$  after similar transformations. These transformations take place in nuclear reactors. The isotopes  ${}_{90}\text{Th}^{232}$  and  ${}_{92}\text{U}^{238}$  used in the process of the manufacture of fissile materials are termed *nuclear raw materials*.

As has been already stated, to obtain a chain reaction the condition  $K_{\text{eff}} \geq 1$  must be fulfilled. The  $K_{\text{eff}}$  depends on the mass of the nuclear fuel. If the mass is small, an appreciable fraction of neutrons leaves it without causing the fission of new nuclei and  $K_{\text{eff}} < 1$ .

Each type of nuclear fuel has its own *critical mass* in which a chain reaction can be maintained ( $K_{\text{eff}} = 1$ ). Thus, for pure uranium  ${}_{92}\text{U}^{235}$  the critical mass is several tens of kilograms. When the mass of a block exceeds the critical mass, an explosion occurs. This is the principle of operation of the atom bomb. It consists of two (or three) blocks of fissile material each with a mass below the critical but with a combined mass exceeding it. To effect the blast the blocks are quickly brought in contact (with the aid of a special fuse). Because of spontaneous fission of the nuclei of fissile material there is always a small number of neutrons in the material to initiate the chain reaction. In the course of an atomic explosion about five per cent of nuclear fuel will be involved in the reaction. Of course, fissile materials should be stored in small blocks at substantial distances from one another.

There are many factors determining the magnitude of the critical mass. Specifically, it depends on the shape: a uranium block in the shape of a sphere has the minimum surface area and therefore the minimum neutron leakage.

Neutron leakage can be reduced by the use of neutron moderators, as well as of shells reflecting neutrons (made, for instance, of beryllium). These methods make it possible to reduce the critical mass of  ${}_{92}\text{U}^{235}$  to a quarter of a kilogram.

#### 42-4 Nuclear Reactors

To maintain the level of nuclear reaction one should be able to exercise a continuous control over the reaction, for even a slight increase in the neutron breeding coefficients to above unity will cause an immediate explosion, while if  $K_{\text{eff}} < 1$  the chain reaction soon dies down.

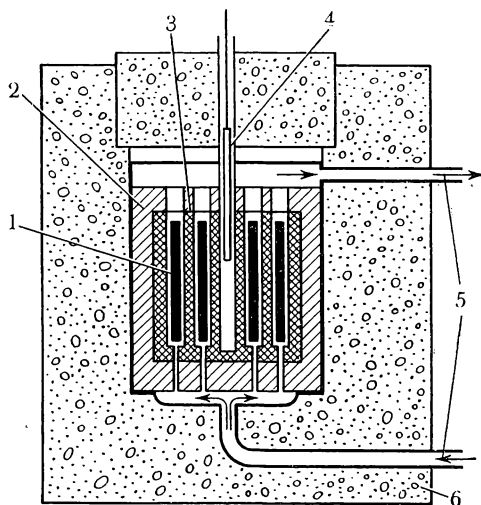


Fig. 42.3 Nuclear reactor: 1, nuclear fuel; 2, neutron reflector; 3, neutron moderator; 4, control rod; 5, heat-transfer agent; 6, protective shell.

This is accomplished in a controlled chain reaction, first realized in 1942 in USA by a team headed by Fermi and independently in the USSR in 1946 by a team headed by Kurchatov. The installation in which the controlled nuclear reaction takes place is called a *nuclear reactor*.

The principal part of the reactor (Fig. 42.3) is the *active zone*, or *core*, in which a self-sustaining chain reaction accompanied by the liberation of energy takes place. To reduce neutron leakage the core is surrounded by a *neutron reflector*.

The reaction is controlled by means of *control rods* made of materials intensively absorbing neutrons (cadmium or boron). With the control rods immersed to a defined depth



into the active zone the reaction proceeds at a constant rate ( $K_{\text{eff}} = 1$ ). This critical mode of operation is constantly maintained with the aid of an automatic device which controls the displacement of the rods and instantly reacts to even the slightest increase or decrease in the reaction rate.

Nuclear reactors using pure fissile materials are feasible, but it is easier and cheaper to use mixtures of isotopes. Often natural uranium in which there is one  ${}_{92}\text{U}^{235}$  atom per 140  ${}_{92}\text{U}^{238}$  atoms or uranium slightly enriched with the  ${}_{92}\text{U}^{235}$  isotope is used as nuclear fuel.

As has been already mentioned, the capture of neutrons by the nuclei of the  ${}_{92}\text{U}^{238}$  isotope is in most cases not followed by fission. Accordingly, a fission chain reaction in a material with a large concentration of  ${}_{92}\text{U}^{238}$  would appear impossible. However, it turns out that  ${}_{92}\text{U}^{238}$  nuclei are very poor absorbers of slow (thermal) neutrons, while the nuclei of the fissile isotope  ${}_{92}\text{U}^{235}$ , on the contrary, absorb slow neutrons much more effectively than fast neutrons. On account of this a chain reaction becomes possible in natural uranium as well if the neutrons generated in the fission process are slowed down (moderated). The uranium itself is a poor moderator because its nuclei are too heavy. The most effective neutron moderators are materials with light atoms. The moderator itself should be a poor absorber of neutrons. Good moderators are helium, which does not absorb neutrons, and heavy water. In practice carbon (in the form of graphite) or ordinary water are used as moderators. The diagram of a chain reaction of slow neutrons is depicted in Fig. 42.4.

The active zone of a slow neutron reactor is filled with a *moderator*, containing rods or plates made of fissile material. The heat liberated in the reaction is carried out of the active zone by a *heat-transfer agent* circulating in special channels. In most cases this function is performed by water circulating at high pressure; gases and liquid sodium are also used. The heat is used to generate steam which drives the turbogenerator of an atomic electric power plant or a propelling plant.

Since the nuclear reactor is a powerful source of highly penetrating neutrons and  $\gamma$ -radiation, it is embedded in a thick *protective shell*.

In a reactor operating on a mixture of  ${}_{92}\text{U}^{235}$  and  ${}_{92}\text{U}^{238}$  isotopes another reaction takes place simultaneously with the chain reaction of fission, that of the transformation of the nuclei of  ${}_{92}\text{U}^{238}$ , which have captured neutrons, into  ${}_{94}\text{Pu}^{239}$  nuclei, that is, nuclear raw material is processed

into fissile material. The plutonium produced in the reactor takes part in the reaction. Thus, the expended nuclear fuel is partly reproduced. The plutonium may also be recovered in a pure form by chemical extraction after the

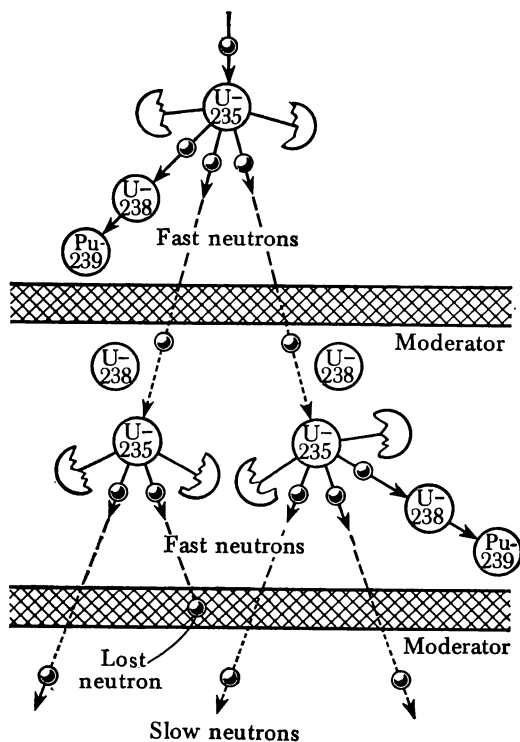


Fig. 42.4 Schematic diagram of chain reaction with neutron moderation.

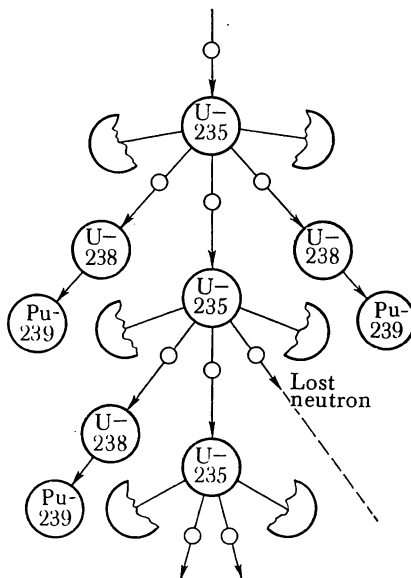
reactor is stopped. Such a method of producing pure fissile material is simpler than the laborious method of uranium isotope separation. The same method is used to produce  ${}_{92}U^{233}$  from thorium  ${}_{90}Th^{232}$ .

If the natural uranium is enriched with the  ${}_{92}U^{235}$  isotope to a concentration of 15 to 20 per cent, the chain reaction becomes possible even without neutron moderation. Such a reactor operating without the moderation of neutrons is termed *fast neutron reactor*, or *breeder reactor*. In the process of its operation more fissile materials are produced than spent (Fig. 42.5). This is because there are no useless losses of neutrons in the moderator; in addition the  ${}_{92}U^{238}$  nuclei which capture fast neutrons experience fission themselves and thus contribute to neutron breeding.

Hence, fast neutron reactors produce energy and do it not only without consuming fissile materials but produce them instead, only nuclear raw materials being consumed.

The active zone of a fast neutron reactor is quite small

Fig. 42.5 Schematic diagram of chain reaction in breeder reactor.



and its cooling is a problem; liquid sodium—the most effective heat-transfer agent but the most inconvenient one because of its chemical activity—has to be used in it.

#### 42-5 Production of Power by Nuclear Reactors

The world consumption of electric power increases rapidly—about twice every decade. To date the principal producers of electric power are thermal power plants burning oil, gas, coal and other kinds of fossil fuel. However, the reserves of such fuels are limited and, moreover, they are valuable for the chemical industry. The most realistic prospect for the replacement of fossil fuel is the use of nuclear fuel. The processing of nuclear fuel—uranium  ${}_{92}\text{U}^{238}$  and thorium  ${}_{90}\text{Th}^{232}$ —promises a multiple increase in nuclear fuel reserves.

The first atomic electric power plant of 5000 kW (5 MW) power was built in the USSR. It went into operation in 1954, that date marking the beginning of the industrial

utilization of nuclear power. In the following years new atomic electric power plants have been built. The largest of them is the Siberian, with a power of 600 MW, and the Novovoronezhsk of 1500 MW power. In the Far North the Bilibinsk atomic electric power plant has been built.

In 1973 the Leningrad plant came into operation. The power output of its first section was 1000 MW (1 million kW). Reactors of equal power are being installed at the Kursk plant. Reactors with a power of 1500 MW have been developed.

The best prospects are offered by fast neutron reactors, which provide for a continuous breeding of fissile materials. In the USSR a breeder reactor with a power output of 350 MW has been developed. Such a reactor operates in the atomic electric power plant in the city of Shevchenko on the Caspian Sea, supplying not only electric power but also fresh water for city use. In sparsely populated regions mobile atomic electric power units of small power are being used.

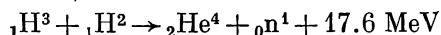
Atomic plants are already producing electric power at a cost comparable to that produced by thermal power plants. A further cut in the costs of nuclear power will bring about a sharp increase in its consumption and by the end of the present century the bulk of electric power will be produced by atomic plants.

Nuclear reactors are also being used for propulsion. In 1959 the Soviet atomic icebreaker *Lenin* was completed. Nuclear-powered submarines make long voyages and can stay under water for a practically indefinite period of time.

## 42-6 Fusion

As was demonstrated in Section 40-11, the average binding energy of a nucleon in the nucleus  $E_b/A$  rises with the increase in  $A$  up to  $A \approx 50-60$  (see Fig. 40.12). Consequently, the fusion of lighter nuclei into a light or medium nucleus should be accompanied by the liberation of energy, because the nucleon bonds in the new nucleus are stronger than in the original nuclei.

The energy liberated is especially high in the case of the fusion of light nuclei because the quantity  $E_b/A$  rises very rapidly at small  $A$ 's. Thus, the energy liberated in the reaction of the fusion of deuterium and tritium into a helium nucleus,



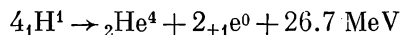
is  $28.3 - (8.5 + 2.2) = 17.6 \text{ MeV}$  (see Table 40.1). The energy produced per each nucleon taking part in the reac-

tion is  $17.6/5 \approx 3.5$  MeV, that is, four times as much as in the reaction of uranium fission. Hence, the energy liberated in the course of the complete transformation of 1 kg of the tritium-deuterium mixture into helium is four times the energy liberated in the complete fission of one kilogram of uranium.

In order to get close enough for the reaction to become possible, the nuclei should possess great kinetic energy, since the coulomb repulsive forces acting between nuclei carrying like charges prevent them from approaching each other.

When a mixture of reacting nuclei is heated to very high temperatures, the kinetic energy of the nuclei becomes high enough for the nuclear fusion reaction, termed *thermonuclear reaction*, to take place.

The necessary conditions exist on the Sun and other stars. The temperature at the centre of the Sun is almost 13 million degrees. At such temperatures the atoms are completely ionized and the substance is in a plasma state consisting only of "bare" nuclei (without electron shells) and of electrons. Inside the Sun a thermonuclear reaction cycle takes place in which hydrogen nuclei are transformed into helium nuclei:

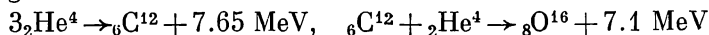


In this cycle an energy almost equal to the binding energy of a helium nucleus  ${}_2\text{He}^4$  is liberated (actually, it is somewhat smaller, because energy is spent to transform two protons into neutrons and positrons).

The approximate composition of the Sun is: hydrogen, about 70 per cent; helium, about 29 per cent; and about one per cent of heavier elements. The mass of the Sun is  $2 \times 10^{30}$  kg. It can be computed that, if the Sun continues to radiate energy at the present rate of  $4 \times 10^{26}$  J/s, the hydrogen will last for  $10^{11}$  years.

From the point of view of its composition and physical properties the Sun is a typical star, the cycle of the thermonuclear reaction of the transmutation of hydrogen into helium being the principal source of energy for the majority of these stars.

Other reactions can also take place inside the stars. As the hydrogen burns out, a helium nucleus is formed at the centre of the star in which the following transmutations can take place at a temperature of about 100 million degrees:



Other thermonuclear reactions also take place.

## 42-7 Controlled Thermonuclear Reaction

The first thermonuclear reactions on the Earth occurred in the explosions of hydrogen bombs. The high temperature required for the thermonuclear reaction was obtained in the course of an atomic explosion.

The hydrogen bomb acts on the following principle. A mixture of deuterium, tritium or other light elements, whose nuclei combine to form helium nuclei, is placed inside a common envelope with an atom bomb. In the course of the explosion of the atom bomb the temperature rises to several tens of millions degrees and a self-sustaining thermonuclear reaction of the transformation of the lighter nuclei into the helium nuclei develops. Thus, the atom bomb serves to "ignite" the mixture of light nuclei.

Since the characteristic parameter of an atom bomb is its critical mass, the power released in an atomic explosion is great but limited. But the mass of light elements in a hydrogen bomb is in principle unlimited. Accordingly, there are no theoretical limits to the power of a hydrogen bomb.

At present work on controlled thermonuclear reaction is in progress in the USSR and other countries. The thermonuclear reaction of small amounts of deuterium (D) and tritium (T) nuclei can be easily realized and is being used in high-voltage D-T tubes to produce neutrons, but the efficiency of this process (the ratio of the energy produced in the reaction to the energy spent to initiate it) is negligible. For the production of energy this ratio needs to exceed unity, that is, the reaction should be self-sustaining. To date it has not been possible to realize a self-sustaining thermonuclear reaction which can be controlled as in nuclear reactors. The crux of the problem is that high temperature is absolutely irrelevant for the fission reaction and is generated only when the heat liberated in the fission reaction is not transferred from the active zone (as in an atomic blast); on the other hand, it is an absolute necessity for a thermonuclear reaction. Accordingly, a very difficult problem has to be solved to make a controlled thermonuclear reaction possible—that of containing high-temperature plasma for a sufficient time.

Inside the stars the plasma is contained by the colossal gravitation pressure of the external layers, its thermal insulation being provided by thick layers separating it from the comparatively cold external layers. Obviously, any walls made of bulk matter are useless because they will instantly evaporate. The main hope appears to be linked with magnetic fields.

High-temperature plasma can be obtained by passing an electric current of a very high density through hydrogen or deuterium. At high-current densities the plasma column is drawn in (pinched) by its own magnetic field to its axis because of the attraction of like currents.

To contain plasma an external magnetic field is also set up around it. Entering this magnetic field, charged particles of plasma acted upon by the Lorentz force move in highly curved paths and are rejected (see Section 25-18). Charged particles are reflected by such concentrated magnetic fields as if from the walls of a vessel, and this earned them the name of *magnetic mirrors*.

These principles are incorporated into the thermonuclear installations, *Tokamaks* (for toroidal chamber with a magnetic field), first developed in the USSR. A Tokamak is, in effect, a transformer with the secondary in the form of a single turn—a chamber in the shape of a torus filled with hydrogen or deuterium. A current of hundreds of thousands of amperes passing through ionized gas converts it to plasma with a temperature of tens of millions of degrees. The magnetic field of this current pinches the ring-shaped plasma column, keeping it from contact with the walls of the chamber. An additional magnetic field produced by coils placed along the torus is also employed to contain the plasma.

A controlled thermonuclear reaction has already been obtained in Tokamaks, but the energy liberated is still less than the energy spent on its initiation, because the temperature and the density of the plasma and the time it is confined are, as yet, too small to involve a substantial number of the nuclei in the thermonuclear reaction.

## 42-8 Some Applications of Radioisotopes

Nuclear physics, in addition to its major technical application in nuclear power production, is being widely applied in many fields of science and technology.

Wide use is made of the high penetration power of  $\gamma$ -radiation, which is many times that of X rays. Since the absorption of radiation increases with the distance it travels through an object, the variations of intensity of radiation can be used to measure the thickness of an object or to detect internal defects. We use  $\gamma$ -radiation to measure small thicknesses.

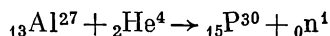
The ionizing capacity of radiation is used to neutralize static charges (for instance, in the textile industry). Friction causes great electrification of the threads (especially the

synthetic ones), with the result that they begin to stick to various parts of the machinery and twisting them becomes difficult. A frequent result is self-ignition. Radiation of radioactive isotopes increases the conductivity of the air and the static charges leak away.

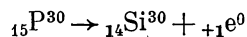
The ionizing effect of radiation is also used in medicine to destroy malignant tumours while  $\gamma$ -radiation kills microbes and is used to sterilize instruments and clothes, to preserve vegetables, fruit, meat, etc.

Absorption of radioactive radiation generates heat that may be used for heating. Such an isotopic heat generator was used for internal heating of the lunar vehicle *Lunokhod-1* during lunar nights.

Most applications of radioactive isotopes are based on the phenomenon of artificial (induced) radioactivity discovered by Frédéric and Irène Joliot-Curie in 1934. They discovered that aluminium, boron and magnesium become radioactive when irradiated with  $\alpha$ -particles. It was established that the irradiation of aluminium with particles induced the nuclear reaction



The phosphorus isotope  ${}_{15}\text{P}^{30}$  is  $\beta^+$ -radioactive and, emitting a positron, turns into a stable silicon isotope:



The discovery of induced radioactivity is a remarkable feat, firstly because for the first time radioactive materials were synthesized, and secondly because it was proved that not only heavy nuclei have radioactive isotopes but light elements as well. For instance, phosphorus has the radio-phosphorus isotope  ${}_{15}\text{P}^{30}$ , and nitrogen the radionitrogen isotope  ${}_{7}\text{N}^{13}$ .

Subsequent research demonstrated that radioactive isotopes of all elements can be synthesized. Most of them emit either  $\beta^-$ -rays or  $\beta^+$ -rays. Radioactive isotopes are produced by irradiating the nuclei with  $\alpha$ -particles, protons, deuterons or high-energy  $\gamma$ -quanta.

Fermi was the first to study radioactivity induced by neutrons absorbed by the nuclei. At present the method of neutron irradiation is the one most widely used for the production of radioactive isotopes. All nuclei except  ${}_2\text{He}^4$  absorb neutrons, usually transforming into  $\beta^-$ -active isotopes. A nuclear reactor is normally used as the source of neutrons.



Products of uranium fission include about 180 radioactive isotopes. Many of them are recovered from the radioactive waste and then put to good use. A list of a few of the newly created radioactive isotopes is given in Table 42.1.

Table 42.1 Some radioactive elements artificially produced

Radioactive element	Half-life period	Radiation emitted
Carbon ${}_6\text{C}^{11}$	20.5 min	positrons
Carbon ${}_6\text{C}^{14}$	5000 yr	electrons
Sodium ${}_{11}\text{Na}^{24}$	14.8 h	electrons, $\gamma$ -rays
Phosphorus ${}_{15}\text{P}^{32}$	14.3 d	electrons
Sulphur ${}_{16}\text{S}^{35}$	87.1 d	electrons
Potassium ${}_{19}\text{K}^{42}$	12.4 h	electrons
Calcium ${}_{20}\text{Ca}^{45}$	152 d	electrons
Cobalt ${}_{27}\text{Co}^{60}$	5.3 yr	electrons, $\gamma$ -rays
Nickel ${}_{28}\text{Ni}^{65}$	160 min	electrons
Bromine ${}_{35}\text{Br}^{82}$	36 h	electrons, $\gamma$ -rays
Strontium ${}_{38}\text{Sr}^{90}$	20 yr	electrons
Silver ${}_{47}\text{Ag}^{106}$	25.5 min	positrons
Iodine ${}_{53}\text{I}^{128}$	25 min	electrons
Cesium ${}_{55}\text{Cs}^{134}$	2.3 yr	electrons, $\gamma$ -rays
Thulium ${}_{69}\text{Tm}^{170}$	129 d	electrons, $\gamma$ -rays
Iridium ${}_{77}\text{Ir}^{192}$	75 d	electrons, $\gamma$ -rays

Artificial radioactive isotopes serve not only as sources of radioactive radiation but are also widely used as *tracers*. Let us dwell on this method of research.

The radioactive isotopes of an element cannot be distinguished by their chemical properties from its stable isotopes. Therefore, by adding a small amount of the radioactive isotopes of the element to a substance one can observe the behaviour of the substance during various processes. By introducing radioactive atoms into a substance we more or less label the molecules which contain them, for they give notice of their presence by radioactive radiation. This is why this method became known as the *method of radioactive tracer atoms*. It features very high sensitivity, since with the aid of a Geiger-Müller counter very small amounts of radioactive atoms can be registered. Let us cite some examples of its application.

Adding some radioactive isotope atoms to a metal and measuring the radioactivity of lubricants, one can determine the rate of wear of surfaces and choose optimal materials

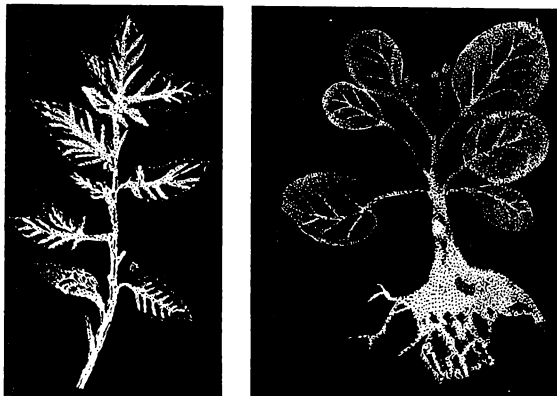


Fig. 42.6 Photographs of plants which assimilated radiophosphorus out of soil. Photographs were produced by radioactive radiation of radiophosphorus.

both for the part and for the lubricants. A method frequently used instead of doping the part with radioactive isotopes in the process of its manufacture is to induce radioactivity in the finished part by irradiating it with neutrons.

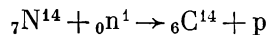
In chemistry the tracer method is used to determine very small solubilities.

Tracer atoms help to estimate the effects on plants of fertilizers, to find out how the most important elements are assimilated. Figure 42.6 shows photographs of plants which have absorbed radioactive phosphorus.

Tracer atoms are used to study photosynthesis in plants. It has been established that in plants oxygen is liberated out of water and not out of carbon dioxide, as was thought before. The tracer atoms method is used to determine the rate of metabolism in the tissues of a living organism; it has been established that tissues are renewed much sooner than it was thought before.

A radioactive "label" makes it possible to observe the circulation of blood in the organism and to detect defects in it. Tracer atoms help to monitor the assimilation of nutrients and medicines and to study the activity of internal organs (for instance, monitoring the accumulation of iodine tracer in the thyroid gland speeds up diagnosis).

An interesting application for radioactive isotopes was found in archeology. In the upper layer of the atmosphere neutrons in the secondary cosmic rays interact with the nuclei of atmospheric nitrogen:



The radioactive carbon  ${}_6\text{C}^{14}$  produced in the reaction is oxidized, mixes with the main mass of atmospheric carbon dioxide, and takes part in the circulation of carbon. The

tissues of plants and animals contain a constant equilibrium concentration of the  ${}_6\text{C}^{14}$  isotope. This concentration declines when the metabolism stops. Knowing the half-life of radiocarbon (5730 years) one can, by measuring the concentration of the remaining radiocarbon in fossils, for instance in the skull of an ancient man, find their age. Radiocarbon has helped provide much valuable information. It was, for instance, established that man appeared in England and America about 10 400 years ago.

The above examples illustrate the wide use of the achievements of nuclear physics in science and technology. However, nuclear power production remains the most important application of nuclear physics. The rapid progress in nuclear technology will solve the most important problem facing mankind—how to satisfy our rapidly growing energy requirements.

The Soviet Union, together with all progressive mankind, is engaged in a relentless struggle for a complete ban on, and the destruction of, nuclear weapons and effectively collaborates with other countries in the peaceful uses of nuclear energy.

part six

# **Astronomy: a Brief Survey**

# The Structure and Evolution of the Universe

## 43-1 The Universe

The study of physical phenomena, processes and regularities on the Earth is closely interconnected with the study of extraterrestrial, astronomical, objects. It suffices to remind the reader that the law of universal gravitation, emission and absorption spectra, and many other laws of nature owe their discovery to astronomical observations. And vice versa, the achievements of “terrestrial” physics and chemistry has facilitated the understanding of what goes on in the space at colossal distances from our planet.

The whole boundless space representing the manifold forms of the existence of matter is termed the *Universe*. The science dealing with the part of Universe accessible to modern means of astronomical observation is termed *cosmology* (from the Greek *kosmos* for Universe and *logos*, for word, speech). Cosmology is divided into the geometry, mechanics and physics of the Universe; its achievements are closely connected with nuclear physics.

The origin and evolution of celestial bodies—planets, stars and galaxies—are studied in *cosmogony* (from the Greek *kosmogonia* for the creation of the world) which is based on the achievements of physics, chemistry, geology and other sciences.

Preceding sections of the course have already dealt with some problems involving the nature of some celestial bodies and methods of studying them. Now let us familiarize ourselves with modern notions about the structure of the Universe.

We live on one of the planets that orbits the Sun due to its gravitational field (Fig. 43.1). The solar system includes, in addition to the Earth and other planets similar

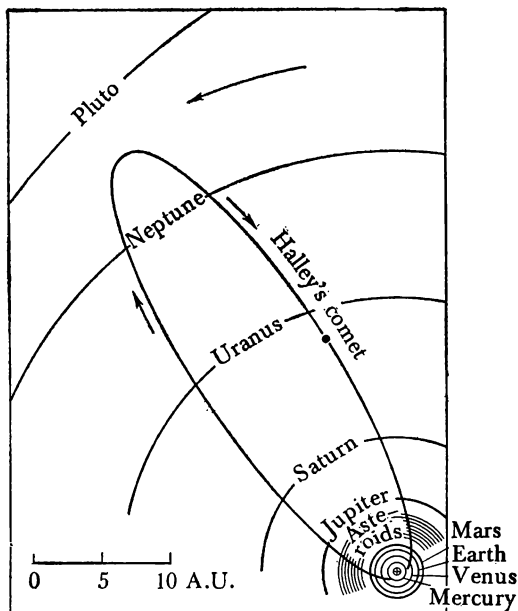


Fig. 43.1 Model of solar system. Scale in astronomical units (1 A.U. =  $1.496 \times 10^{11}$  m). The dimensions of orbits of planets of Earth's group have been enlarged for better visibility.

to it (Mercury, Venus and Mars), the giant planets (Jupiter, Saturn, Uranus and Neptune) and the little studied planet Pluto. The enumerated planets, the principal information on which is tabulated in Table 43.1, are termed *big planets*. In addition to them other planets known to date include about 2000 *minor planets* (*asteroids*) whose diameter is much smaller than that of the big planets (the diameter of the smallest asteroid is of the order of 1 km). Bodies of smaller dimensions are not visible in a telescope and we learn about them only when, upon meeting our Earth, they fall on its surface as meteorites. These are bodies consisting of iron and silicates which move in the solar system mainly between the orbits of Mars and of Jupiter and are like asteroids.

Another class of minor bodies of the solar system is made up of numerous comets (one of them is illustrated in Fig. 43.1). A small number of them remain all the time inside planetary orbits and revolve about the Sun with periods from several years to several decades. Most comets, however, move outside planetary orbits; but sometimes moving in extremely elongated elliptical orbits they come close to the Sun. At this stage their nuclei, which consist of frozen gases (methane and ammonia) and ice with high melting point particles frozen in it, exude these gases. Together with dust released by the nucleus, they form the head and the tail of the comet, the latter extending for tens of millions of kilometres. Despite its colossal dimensions the density of substance in the tail is quite negligible (it should be noted that the total mass of a comet does not exceed a billionth ( $10^{-9}$ ) fraction of the mass of the Earth). The luminosity of the comets is due to the fluorescence of the gas (see Section 38-17) excited by solar radiation and to the scattering of sunlight by the dust.

Table 43.1 Solar system

	Sidereal period, years	Average distance from Sun, $10^6$ km	Mass (Earth = 1)	Density, $\text{kg/m}^3$	Equatorial diameter, km	Sidereal period of rotation about axis	Number of satel- lites
Mercury	0.241	58	0.05	5600	4900	58.65 days	
Venus	0.615	108	0.81	5200	12100	243.0 days	
Earth	1.000	150	1.00	5500	12756	23 h 56 min 4 s	1
Mars	1.881	228	0.11	4000	6800	24 h 37 min 23 s	2
Jupiter	11.86	778	348	1300	142000	9 h 50 min	12
Saturn	29.46	1426	95.1	700	120000	10 h 14 min	10
Uranus	84.01	2869	14.5	1500	50000	10.8 h	5
Neptune	164.7	4496	17.3	1700	50000	15.8 h	2
Pluto	248.9	5929	?	?	2000 (?)	6.4 days	?
Sun			333000	1400	1392000	25.5 days	

The solid substance contained in the planets, the gases constituting their atmospheres, and comet tails together with minor bodies and cosmic dust make up only  $1/750$  of the total mass of the solar system. The main mass of the solar system is concentrated in its central body, the Sun.

At this stage of evolution of the Universe the predominant part of the substance contained in it exists in the form of blobs of hot plasma like the Sun and the stars. The stars are separated by vast distances. Thus, the star closest to the Sun is at a distance 270 000 times further from it than

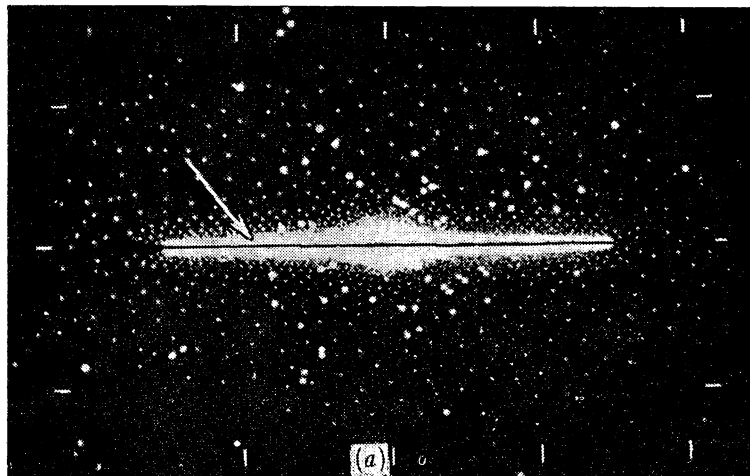
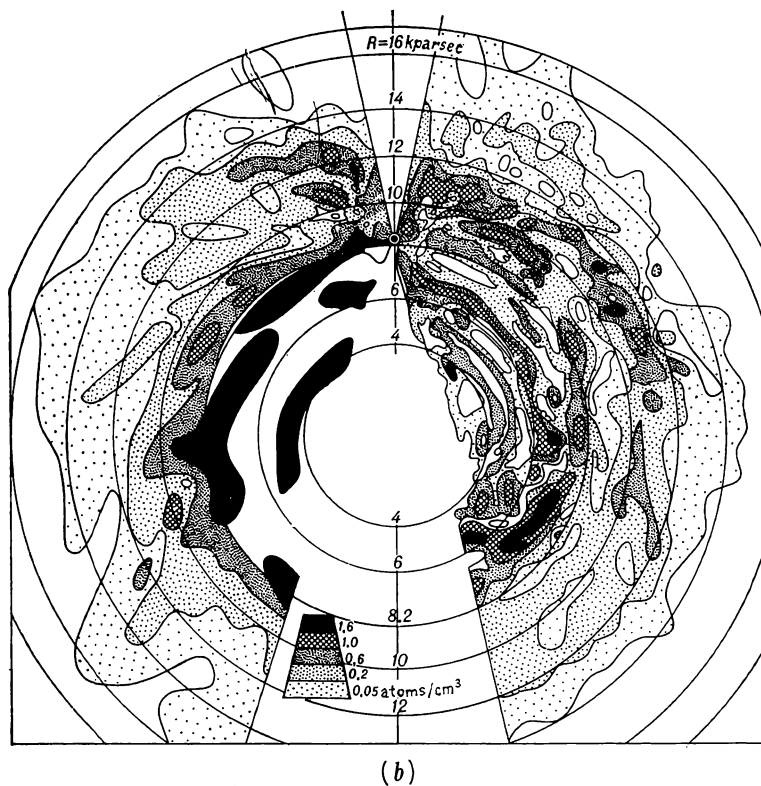


Fig. 43.2 Schematic diagram of our Galaxy: (a) in plane normal to that of Milky Way; (b) in plane of Milky Way.





the Sun is from the Earth, that is, the separation between the stars in the vicinity of the Sun is approximately ten million times greater than their own dimensions.

Together with other stars the Sun constitutes a colossal stellar island, the *Galaxy*. The Galaxy is a concentration of approximately 150 billion stars and interstellar matter. We observe our Galaxy in the form of a bright band crossing the sky (it has been called the Milky Way). About 98 per cent of the galactic mass is concentrated in the stars, only per cent being reserved for the interstellar substance (gas together with dust in a proportion of  $100 \div 1$ ). The average concentration of interstellar substance is about 1 particle per  $1 \text{ cm}^3$ , but sometimes interstellar substance is found in the form of denser clouds.

The diameter of the Galaxy is about 30 thousand parsecs (pc) (see Section 1-6), the thickness of the disk in which the majority of the stars is concentrated being 460 pc. Figure 43.2 is a schematic diagram of the Galaxy, the arrow indicating the position of the Sun.

The distribution of stars of various types and of other objects throughout the Galaxy is not the same. For instance, the hottest and more massive stars together with the gas nebulae are concentrated near the plane passing through the centre of the Galaxy (the *planar system*). The Sun, too, is close to this plane. It is because of this concentration that we observe the Milky Way as such. Other types of stars and spherical star concentrations are scattered throughout the Galaxy, but for the most part such objects are situated in the vicinity of its centre (the *spherical system*). It will be demonstrated below that these types of distribution, as well as other types, are conditioned by the different ages of the stars.

The stars and other objects in the plane of the Milky Way (Fig. 43.2*b*) form spiral branches diverging from the centre of the Galaxy. These spiral branches are made easily visible not so much by the stars as by the interstellar gas, the distribution of which is studied by radioastronomical methods. All objects constituting the Galaxy revolve about an axis passing through its centre and are held together by gravitation. Our Sun takes 200 million years to complete one revolution.

Another galaxy similar to ours is about 0.55 million parsecs (approximately 20 times the dimensions of our Galaxy) away from us. This is the Andromeda Galaxy (Fig. 43.3), visible to the naked eye. At present we are able to study galaxies several billion light years away from

us. Photographs obtained with the most powerful telescopes depict hundreds of millions of galaxies.

Studying distant galaxies we see the past of the Universe, for the light reaching us was emitted several billion years ago. Some galaxies, which because of the vast distances separating them from us are seen as small bright spots,

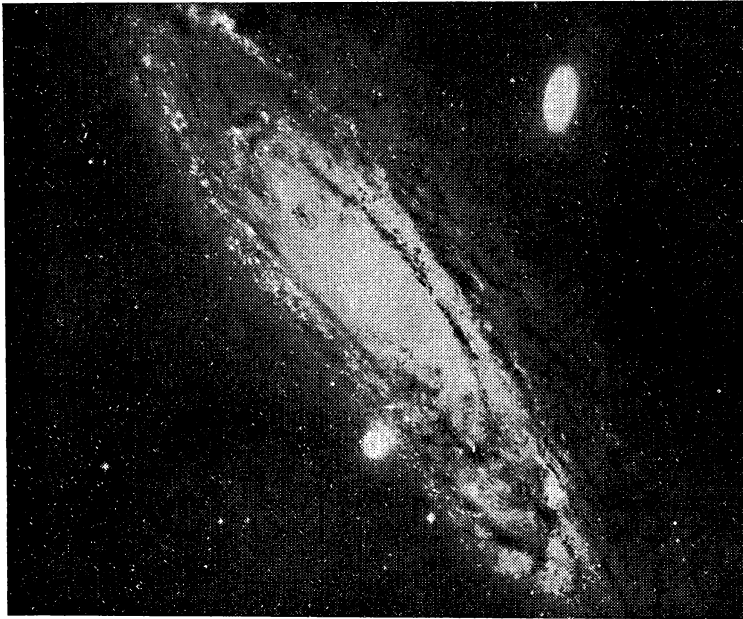


Fig. 43.3 Andromeda Galaxy—galaxy nearest to us and in many ways similar to ours.

radiate enormous amounts of energy in the radiofrequency range and for this reason are termed *radiogalaxies*. Studying their optical radiation one is bound to observe that the spectra of the majority of galaxies are absorption spectra characteristic of the stars which constitute them. The absorption lines in the spectra of all galaxies (except those nearest to us) are displaced in the direction of longer waves (see Section 37-14). The magnitude of this *red shift* is directly proportional to the distance to the galaxy. This means that the galaxies recede at speeds increasing with the distances between them. The causes for this recession of the galaxy system as a whole will be discussed in Section 43-3. The red shift shows that the Universe cannot be regarded as stationary. It is the scene of the continuous evolution of individual celestial bodies and of the Universe itself.

The distribution of galaxies in space is nonuniform. Like the stars they form separate groups and clusters. For instance, our Galaxy together with neighbouring galaxies forms a local system—a group consisting of about 20 objects. This group, in turn, is part of a large cluster made up of several thousand galaxies. The difference between constellations and galaxy clusters is that the distances separating the latter are only several times greater than the dimensions of the galaxies themselves. Accordingly, as the scale is increased one becomes entitled to speak of a uniform distribution of substance in the Universe.

### 43-2 The Origin and Evolution of Celestial Bodies

One of the greatest achievements of the twentieth-century astronomy was that it established that the process of star formation is a continuous one (taking place in our time as well). It has been established that many of the visible stars are younger than our planet and some of them were born quite recently, when man already existed on the Earth.

The majority of scientists believe that stars are formed by the condensation of clouds of rarefied interstellar matter consisting of gases and dust which, acted upon by gravitational forces, form a dense nontransparent gaseous sphere. In the early stages the gaseous pressure inside this comparatively cold sphere is incapable of withstanding the gravitational forces which continue to contract it. But as the contraction continues the temperature in the interior of the star rises to a level sufficient for a thermonuclear reaction (see Section 42-6). At this stage the pressure of the hot gas inside the future star matches the gravitational forces and the contraction stops. The process described takes a comparatively short time, from several million to several hundred million years (depending on the mass of the star).

The stars radiate energy generated by thermonuclear reactions taking place in their central parts. The life of the star in this stage is also determined by its mass. It is the star's mass which determines the consumption rate of its reserves of hydrogen as it transforms into helium. Thus, hot giant stars with masses from 10 to 20 times the mass of our Sun will use up their nuclear fuel in several million years, while our Sun and other stars with a similar mass produce stable radiation for 10 to 15 billion years.

Still, a time comes when there is no more hydrogen left in the nucleus of the star, no more energy is released, and

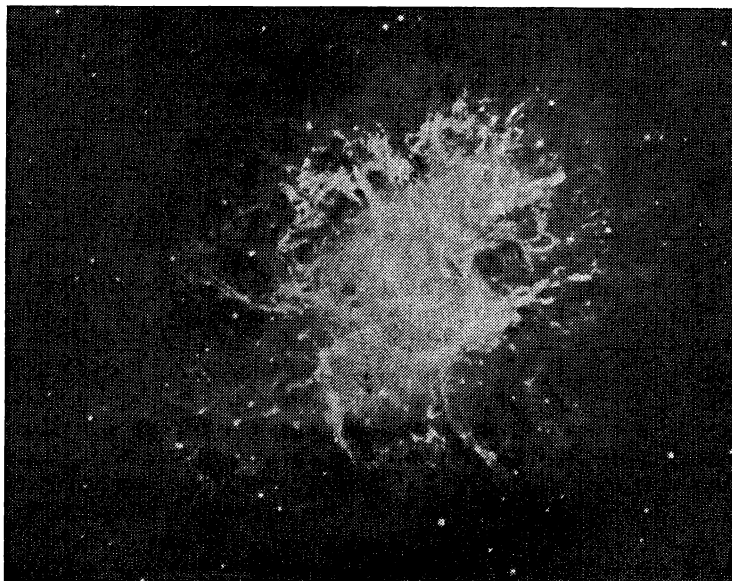
gravitational forces begin to contract the nucleus. At this stage thermonuclear reactions can take place only inside a comparatively narrow layer surrounding the nucleus. At this stage the luminosity\* of the star and its size should increase. The evolution of the star is accelerated and it turns into a *red giant*. When the temperature of the contracting helium nucleus reaches 100-150 million degrees, a new type of reaction sets in: a carbon nucleus is synthesized from three helium nuclei (see Section 42-6). Calculations predict that our Sun will turn into a red giant in 8 billion years and will remain at this stage for several hundred million years, its luminosity being several hundred times and its radius several tens of times as great as at present.

Such giant stars soon use up their nuclear fuel reserves and lose a greater part of their mass, either gradually or by rejecting their outer shell. In the final stages of evolution stars with masses comparable to that of the Sun turn into *white dwarfs*. At this stage only the central dense part in which nuclear reactions have already ceased is left of the star. Such stars gradually cool down, their radiation diminishes and they become invisible. Their dimensions are smaller than the Earth's, but since their masses are comparable to the mass of the Sun the density of substance in them is millions of times greater than the density of water.

However, not all stars are destined to follow such a relatively calm course of evolution. Some of them experience catastrophic changes in the process of evolution. The term used for such cases is *supernova* and it results in very substantial changes in the star's structure. Characteristic nebulae have been detected where supernovae occur (Fig. 43.4), all of them without exception being powerful sources of radiofrequency radiation (see Section 37-17). The mass of the gases ejected in the more powerful outburst may exceed by several times the mass of the Sun. If the part of the star remaining after the outburst has a mass of about 1.5 times the mass of the Sun it cannot turn into a white dwarf. The forces of gravity contract it to a much smaller size. The diameter of such objects is of the order of 10 km and their average density is about  $10^{18}$  kg/m<sup>3</sup>, greater than that of an atomic nucleus. Such stars have received the name of *neutron stars*, because at such densities the substance consists entirely of neutrons formed as the result of the fusion of protons and electrons.

\* The luminosity of a star characterizes the energy flux radiated per unit time.

Fig. 43.4 Crab Nebula—  
remnants of supernova.



The theory of such stars was evolved as far back as the thirties by the Soviet physicist Lev D. Landau (1908-1968), but they were found only in 1967 as sources of radio-frequency radiation emitting strictly periodic short pulses (of the order of a second or fraction of a second). The cause of the strict periodicity of the radiofrequency pulses and of the optical pulses emitted by *pulsars* is their rapid rotation. The shortest period of radiofrequency and optical pulses is that of a pulsar discovered in the well-known *Crab Nebula* and at the location of the supernova of 1054. Its period is 0.033 s.

Still more wonderful objects should come into being in the final stages of the evolution of a star if its mass after its nuclear fuel reserves have been exhausted exceeds 1.6 of the mass of the Sun. In this case the pressure of the so-called degenerate gas which makes up the star in the last stage of its evolution is no longer able to stop its contraction by gravitational forces. The star's density will grow as it contracts at enormous speed. Its mass will remain constant, but the speed  $v_{\text{par}}$  a body needs to have in order to leave its gravitational field (the so-called *parabolic*, or *escape, speed*) will grow. After the object reaches the radius for which  $v_{\text{par}} \approx c$ , neither particles nor radiation will be able to escape from its gravitational field. For this reason such objects have been given the name *black holes*. They

are invisible but interact with the outside world through gravitational forces. A black hole can be detected, for instance, if it is one of two binary stars bound by gravitational forces and revolving about a common centre of mass; in this case the existence of the black hole can be inferred from the behaviour of the second normal visible star. Scientists believe that one of the components of the binary star Cygnus-X is a black hole and hope to discover other similar objects.

Now let us turn to the origin of planets. Although planetary systems are peculiar not only to the Sun but to other stars as well, at present the planets of other stars cannot be observed even with the best existing telescopes. Therefore all conclusions about the origin and evolution of the planets have to be made on the basis of information gained in the study of only one example, that of the solar system.

All modern hypotheses about the origin of the Earth and the planets are based on the idea of their formation from a gas-dust cloud, most scientists favouring the idea that the formation of the Sun and the planets from this cloud took place simultaneously. The cloud's composition was close to that of the contemporary composition of the Sun, 98 per cent being hydrogen and helium and 2 per cent other elements which formed various compounds and condensed into particles.

The dust gradually assembled in a single plane, forming a layer of increased density. This layer did not remain uniform but disintegrated into separate blobs which, colliding with each other, associated and contracted. The solid bodies formed as a result also experienced collisions and either split or grew assimilating the fragmentary substance. Eventually only nine nuclei grew to large sizes and became the big planets. This idea of the planets being formed in the process of the association of solid bodies and dust was advanced by the eminent Soviet scientist Otto Yu. Shmidt (1891-1956). It brought about a true revolution in the cosmogony of planets, taking the place of the former notions about the condensation of planets from gaseous accumulations. The American chemist Harold C. Urey (b. 1893) independently found support for this idea from his physicochemical studies of the composition and structure of meteorites.

The gas component of the preplanetary cloud was greatly affected by the solar wind—a powerful stream of particles emitted by the Sun, formerly in even greater quantities than now. The planets similar to the Earth and formed in the vicinity of the Sun consist mainly of silicates and metals. At great distances from the Sun, where Jupiter

and Saturn were formed, there still remained a substantial mass of gases (of hydrogen and of helium) and they were drawn into the planets. Thus, Shmidt's hypothesis explains why the planets are subdivided into two groups on account of their physical properties.

Let us now study the evolution of galaxies. Our Galaxy and other galaxies which are, in effect, great accumulations of stars and interstellar matter, together with all the bodies which make them up, are subject to substantial changes in the course of time.

Firstly, if one takes into account all that has been said above about the origin and the evolution of the stars, one is bound to see that the amount of interstellar matter gradually diminishes.

Secondly, during the time this matter exists in the form of stars it changes its chemical composition: the amount of hydrogen decreases and the amount of helium and other elements formed as a result of thermonuclear processes increases at the expense of the former. The heaviest elements are synthesized only in special condition, in the course of the supernova catastrophes. Hence, the following generation of stars is formed from a substance of new chemical composition.

Observing the distribution of the stars of different composition one is able to study the distribution of stars in a galaxy or in a star cluster by their age. It has been established that the oldest objects in a galaxy form a spherical system. This means that the gas cloud from which the galaxy was formed was of a spherical shape. The mass of the gas contracted and flattened, assembling in a plane normal to the galaxy's axis of rotation. The subsequent process of star formation took place in a disk close to this plane. A further flattening of the disk was opposed by the magnetic field whose lines of force determine the spiral shape of the distribution of interstellar hydrogen and of the stars in the galactic disk formed from it.

The spiral branches of a galaxy are in some way or other connected with its *nucleus*. The nuclei of the galaxies, their central parts, are not just regions of higher star distribution density. In recent years numerous facts have been discovered pointing to a high level of activity of galactic nuclei. The first one to note the peculiarities of galactic nuclei was the eminent Soviet astrophysicist Viktor A. Ambartsumyan (b. 1908). Observations carried out in a wide spectral range from radiowaves to X rays have demonstrated that the radiation power of the galactic nuclei changes noticeably over a period of several months or even weeks.

Calculations show this to be the result of processes taking place inside small volumes. These processes are accompanied by the liberation of energy much greater than that liberated in the most powerful star explosions. Especially manifest is the activity of the galactic nuclei in the radiofrequency band and, accordingly, they have been termed *radiogalaxies*. Other types of galaxies with active nuclei have also been observed. There is ample ground to suppose that the nucleus of our Galaxy was formerly very active.

An especially intensive activity is observed in *exploding galaxies*. One such galaxy is illustrated in Fig. 43.5. The gas ejected out of the central cluster several million years ago disperses at a speed of about 1000 km/s; the combined mass of ejected gas is  $5 \times 10^9$  times that of the Sun.

Radioastronomical observations led to the discovery in 1963 of wonderful objects, *quasars*. This is an abbreviation for *quasi-stellar radio sources*. By their external appearance they cannot practically be distinguished from the stars, but radio observations show a very complex structure,



Fig. 43.5 Active galaxy with gas ejected from its nucleus.

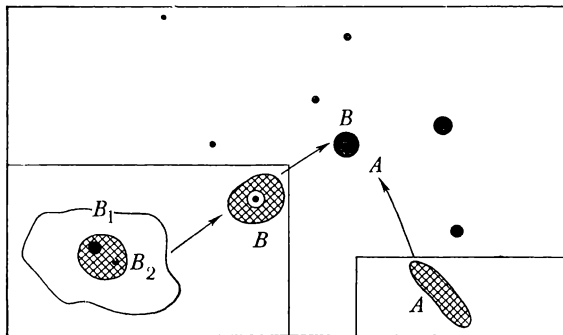


Fig. 43.6 Structure of quasar established by radio observations. Plasma blobs are visible in detail on enlarged scale below. Length of protuberance is about  $3 \times 10^4$  pc.

the presence in the quasars of compact objects of spherical shape and various protuberances and blobs of different sizes (Fig. 43.6). Apparently, quasars, like galaxies, cannot consist of individual stars. They are similar to the galactic nuclei, but radiate still more energy. A quasar of a diameter not greater than a hundredth of a parsec can radiate energy many hundreds of times greater than the energy radiated by such a galaxy as ours. Apparently, quasars are rotating bodies consisting of plasma and strong magnetic fields.

### 43-3 Cosmology

Cosmology studies the structure of the Universe as a whole and its evolution in time. The concept of an essen-



tially unchanging (stationary) Universe leads to some paradoxical conclusions not agreeing with observations.

Firstly, it turns out that the force of interaction of a body with all the masses in a homogeneous stationary Universe is indefinite. Secondly, if one assumes the Universe to be stationary, one cannot explain such a fundamental experimental fact as the red shift in the galactic spectra. As has been already mentioned above, the recession speed of the galaxies  $v$  is related to the distance  $R$  by a linear expression

$$v = H \times R$$

the term for it being *Hubble's law* named after the American scientist Edwin P. Hubble (1889-1953) who in 1929 proved the existence of such a regularity. The term for the factor  $H$  which is at present rated at 50-100 km/(s · Mpc) is *Hubble's constant*. It shows that the recession speed of galaxies increases by 50-100 km/s with an increase in their distance of 1 Mpc (1 Mpc =  $10^6$  pc).

The nonstationary state of the Universe is at present beyond all doubt. The cause for it may be understood already in the context of Newtonian mechanics; this is renewed proof of the continuity existing between the former and the theory of relativity, which throws light on more intimate regularities of nature but does not disavow Newtonian mechanics.

If one imagines all the known galaxies occupying the space of a gigantic sphere and interacting in accordance with the law of universal gravitation, it becomes clear that such an accumulation of mutually attractive objects cannot remain stationary. If initially those objects were at rest, the mutual attraction will cause them to approach and the radius of the sphere will diminish. On the other hand, if their initial speeds were directed away from the sphere's centre, their subsequent motion would depend on the initial speeds. For large enough speeds the radius of the sphere would grow indefinitely. Small outward speeds would entail a change from an increasing to a decreasing spherical radius. The magnitude of the initial speed itself, which determines the nature of the expansion, depends in the long run on the density of the sphere. The theory of relativity only refines the conclusion about the nonstationary state of the Universe.

Both observation and theory uphold the notion of a Universe in a state of evolution. For example, it follows from the spectra of quasars that they recede from our Galaxy at enormous speeds, in some cases as high as 100-200 thousand km/s, that is, comparable to the speed of

light. Hence, they are among the most distant objects in the Universe. Thus, we are able to see the past of the Universe and that several billion years ago the substance in the Universe was in a state different from today's.

The quantity reciprocal to the Hubble's constant ( $1/H \approx \approx 10^{10}$  years) is often called the *age of the Universe*. This is taken to mean that if the recession of the galaxies always obeyed the same law,  $10^{10}$  years ago all of them would have been concentrated in a very small space. Accordingly, the substance should have been in a state of high density and high temperature. This hypothesis of a hot Universe advanced 25 years ago is confirmed by observations.

Progress in radioastronomy made it possible to discover in 1965 a radiation with maximum spectral density in the millimetre waveband. Apparently, it is not connected with any bodies at present existing in the Universe and reflects the distribution of its substance in the past, in the initial stages of its evolution. This radiation, termed *relict radiation*, provides information on the state of the Universe in an age when its dimensions were much smaller than at present.

Thus, astrophysics poses problems relating to the behaviour of matter at temperatures of the order of several billion degrees and having densities comparable with those of atomic nuclei. Diving still deeper into the past, cosmology approaches the problems of the state of matter in the early stages of evolution and the "beginning of time".

The concepts of space and time are inseparable from the concept of matter, the interconnection of those concepts being much more pronounced in such extreme conditions than in normal conditions. According to the general theory of relativity the geometry of space-time depends on the distribution of matter in the Universe. This dependence finds its expression, specifically, in the dependence of the nature of universal recession on average density of matter.

The corollaries of the general theory of relativity (of the theory of gravitation) are now being checked in astronomy and in cosmology, which deal with distances of the order of several billion light years ( $10^{26}$  m). One of the most important corollaries of the theory is that the deflection of light in the vicinity of massive bodies could be observed at smaller distances. This phenomenon has been observed for the light coming from the stars which form the background for the Sun in times of solar eclipses.

The cosmological problem—the problem of the structure and the evolution of the Universe as a whole—is one of the most fundamental problems of modern science.

# APPENDIX

## Names, Symbols and Conversion Equivalents of SI Units

Quantity	Name and symbol	Conversion equivalents
BASE UNITS		
length	metre (m)	1 centimetre (cm) = $10^{-2}$ m 1 kilometre (km) = $10^3$ m
mass	kilogram (kg)	1 gram (g) = $10^{-3}$ kg 1 quintal (q) = $10^2$ kg 1 tonne (t) = $10^3$ kg (metric ton)
time	second (s)	1 minute (min) = 60 s 1 hour (h) = 3600 s
electric current	ampere (A)	1 abampere (abA) = 10 A 1 statampere (statA) = $1/3 \times 10^{-9}$ A
thermodynamic temperature	kelvin (K)	$K = ^\circ C + 273.15$
luminous intensity	candela (cd)	
amount of substance	mole (mol)	
SUPPLEMENTARY UNITS		
plane angle	radian (rad)	1 degree (deg) = $\pi/180$ rad 1 minute (min) = $\pi/108 \times 10^{-2}$ rad 1 second (sec) = $\pi/648 \times 10^{-3}$ rad
solid angle	steradian (sr)	

## Continuation

Quantity	Name and symbol	Conversion equivalents
DERIVED UNITS, MECHANICS		
area	square metre (m <sup>2</sup> )	1 square centimetre (cm <sup>2</sup> ) = 10 <sup>-4</sup> m <sup>2</sup> 1 square kilometre (km <sup>2</sup> ) = 10 <sup>6</sup> m <sup>2</sup>
volume	cubic metre (m <sup>3</sup> )	1 cubic centimetre (cm <sup>3</sup> ) = 10 <sup>-6</sup> m <sup>3</sup> 1 litre (l.) = 1.000 000 × 10 <sup>-3</sup> m <sup>3</sup>
density	kilogram per cubic metre (kg/m <sup>3</sup> )	1 gram per cubic centimetre (g/cm <sup>3</sup> ) = 10 <sup>3</sup> kg/m <sup>3</sup>
velocity	metre per second (m/s)	1 centimetre per second (cm/s) = 10 <sup>-2</sup> m/s 1 kilometre per hour (km/h) = 1/3.6 m/s
acceleration	metre per second squared (m/s <sup>2</sup> )	1 centimetre per second squared (cm/s <sup>2</sup> ) = 10 <sup>-2</sup> m/s <sup>2</sup>
force	newton (N)	1 kilogram-force (kgf) = 9.806 65 N 1 dyne (dyn) = 10 <sup>-5</sup> N
work, energy	joule (J)	1 kilogram-force-metre (kgf·m) = 9.806 65 J 1 erg (erg) = 10 <sup>-7</sup> J
power	watt (W)	1 kilogram-force-metre per second (kgf·m/s) = 9.806 65 W 1 horsepower (hp) = 735.499 W 1 erg per second (erg/s) = 10 <sup>-7</sup> W
pressure	pascal (Pa)	1 dyne per centimetre squared = 0.1 Pa 1 mm mercury (mmHg) = 133.322 Pa 1 technical atmosphere (at) = 1 kgf/cm <sup>2</sup> = 98 066.5 Pa 1 physical atmosphere (atm) = 101 325 Pa
angular velocity	radian per second (rad/s)	
angular acceleration	radian per second squared (rad/s <sup>2</sup> )	
frequency	hertz (Hz, s <sup>-1</sup> )	

## DERIVED UNITS, HEAT

quantity of heat	joule (J)	1 calorie (cal) = 4.186 J 1 kilocalorie (kcal) = 4186 J
specific heat capacity	joule per kilogram kelvin J/(kg·K)	1 kilocalorie per kilogram degree Celsius (kcal/(kg·°C)) = 4186 J/(kg·K)
specific latent heat of change of phase	joule per kilogram (J/kg)	1 kilocalorie per kilogram (kcal/kg) = 4186 J/kg

Quantity	Name and symbol	Conversion equivalents
DERIVED UNITS, ELECTRICITY, MAGNETISM, OPTICS		
quantity of electricity (electric charge)	coulomb (C)	1 abcoulomb (abC) = 10 C 1 statcoulomb (statC) = $1/3 \times 10^{-9}$ C
electric field strength	volt per metre (V/m)	1 statvolt per centimetre (statV/cm) = $3 \times 10^4$ V/m
potential, potential difference, electromotive force	volt (V)	1 abvolt (abV) = $10^{-8}$ V 1 statvolt (statV) = 300 V
capacitance	farad (F)	1 abfarad (abF) = $10^9$ F 1 statfarad (statF) = $1/9 \times 10^{-11}$ F (also known as centimetre)
electric conductance	siemens (S, mho, $\Omega^{-1}$ )	
electric resistance	ohm ( $\Omega$ )	1 abohm (ab $\Omega$ ) = $10^{-9}$ $\Omega$ 1 statohm (stat $\Omega$ ) = $9 \times 10^{11}$ $\Omega$
work, energy	joule (J)	1 kilowatt-hour (kWh) = $3.6 \times 10^{11}$ J 1 electronvolt (eV) = $1.602\,192 \times 10^{-19}$ J
power, electrical output	watt (W)	1 horsepower (hp) = 735.499 W
magnetic flux	weber (Wb, V·s)	1 maxwell (Mx) = $10^{-8}$ Wb
magnetic flux density (magnetic induction)	tesla (T, Wb/m <sup>2</sup> )	1 gauss (G) = $10^{-4}$ T
magnetic field strength	ampere per metre (A/m)	1 oersted (Oe) = $10^3/4\pi$ A/m
inductance	henry (H, V·s/A)	1 abhenry (abH) = $10^{-9}$ H (also known as centimetre) 1 stathenry (statH) = $9 \times 10^{11}$ H
luminous flux	lumen (lm, cd·sr)	
luminance	nit (nt, cd/m <sup>2</sup> )	1 stilb (sb) = 1 cd/cm <sup>2</sup>
illuminance	lux (lx, lm/m <sup>2</sup> )	

# NAME INDEX

Abelson, P. H., 600  
Ambartsumian, Victor A., 628  
Anderson, Carl D., 587  
Aston, Francis W., 576

Balmer, J. J., 526  
Basov, Nicolai G., 534  
Becker, Howard S., 573  
Becquerel, Antoine Henri, 567  
Biot, Jean Baptiste, 295  
Blackett, Patric M. S., 572  
Bohr, Niels, 526  
Boltzmann, Ludwig, 67  
Bothe, Walther, 573  
Boyle, Robert, 74  
Brown, Robert, 49  
Bunsen, W. E., 480  
Butler, Clifford C., 591

Carnot, Sadi, 84  
Chadwick, Sir James, 572  
Charles, Jaques A. C., 63  
Cherenkov, Pavel A., 571  
Coulomb, Charles A., 175  
Cowan, C. L., 590  
Curie, Marie, 567  
Curie, Pierre, 567

Dalton, John, 36, 109  
Dirac, Paul A. M., 588  
Doppler, Christian J., 496

Einstein, Albert, 513  
Euclid, 403

Faraday, Michael, 114  
Fermi, Enrico, 584  
Foucault, Jean Bernard, 320, 337  
Fraunhofer, Joseph, 493  
Frisch, Otto R., 601

Gay-Lussac, J. L., 73  
Geiger, Hans, 564

Hahn, Otto, 601  
Heisenberg, Werner Karl, 574  
Helmholtz, Hermann von, 95  
Hertz, H., 381, 384  
Hooke, Robert, 149  
Hubble, Edwin P., 630  
Huygens, Christian, 337, 394

Ivanenko, Dmitri D., 574

Joliot-Curie, Frédéric, 613  
Irène, 613  
Joule, James P., 31, 334

Kapitza, Peter L., 117, 160  
Kelvin, Lord, 65  
Kirchhoff, Gustav R., 486  
Kurchatov, Igor V., 601

Landau, Lev D., 161, 626  
Laplace, Pierre S., 130  
Lane, Max von, 600  
Lawrence, Robert O., 598  
Lebedev, P. N., 396, 505  
Lenin, V. I., 19  
Lenz, Heinrich P. E., 234  
Lodge, Sir Oliver J., 364  
Lomonosov, M. V., 84  
Lorentz, Hendrick Antoon, 305  
Loschmidt, N., 55

Marconi, Guglielmo, 301  
Mariotte, Edmé, 74  
Maxwell, James Clerk, 51, 52, 380  
Mayer, Julius K. von, 95  
McMillan, E. M., 600  
Meitner, Lise, 601  
Mendeleev, Dmitri, 114  
Michelson, Albert A., 399  
Millikan, Robert A., 399  
Morley, Edward W., 543  
Müller W., 564

Newton, Sir Isaac, 476

Oersted, Hans Christian, 283  
Ohm, Georg Simon, 218

Pauli, Wolfgang, 589  
Peltier, Jean C. A., 242  
Petrov, V. V., 259  
Planck, Max, 396, 491  
Poiseuille, Jean L. M., 136  
Popov, Alexander S., 381  
Powell, Cecil Frank, 581  
Prokhorov, Alexander M., 534

Reines, F., 590  
Rochester, George D., 591  
Roemer, Ole, 399  
Roentgen, Conrad Wilhelm, 498  
Rumford, Count, 84  
Rutherford, Lord Ernest, 170

Savart, Felix, 295  
Schmidt, Otto Yu., 627  
Soddy, Frederick, 567, 576  
Stern, Otto, 50  
Stoletov, Alexander G., 510  
Stokes Sir George G., 531  
Stoney, George J., 170  
Strassmann, Fritz, 601

Thomson, Benjamin, 84  
Thomson, Sir Joseph John, 576  
Thomson, Sir William, 65

Trey, Harold C., 627

Vavilov, Sergei L., 571  
Volta, Alessandro, 251

Wien, Wilhelm, 492  
Wilson, Charles Thomas R., 565

# SUBJECT INDEX

- Aberration, chromatic, 432
  - spherical, 431
- Absolute zero, 64
- Absorptance, 488
- Absorption of sound, 359
- Acceptor, 275
  - conductivity, 275
- Accommodation, 435
- Accumulator, 216, 253
- Activators, 532
- Acuity, visual, 438
- Adaptation of eye, 436
- Adsorption, 110
- Aerial, 381
- Air, 118
- Alloys, 156
- Alpha-rays, 170
- Ammeter, 214, 304
- Ampere, 290
- Ampere force, 290
- Ampere-hour, 254
- Ampereturns, 296
- Amplitude, 328
  - of e.m.f., 363
- Amorphous substance, 137
- Analysis, spectral, 494
- Analyzer, 461
- Angle, contact, 129
  - critical, 412
  - deflection, 413
  - of incidence, 346
  - phase, 332
  - of reflection, 346
  - refraction, 401, 413
  - of view, 437
- Anisotropy, 139
- Annihilation of particles, 559
- Anode, 238
  - diode, 238
- Anode dissolution, 246
  - reaction, 245
- Antenna, 381
- Antimatter, 598
- Antineutrino, 589
- Antinodes, 348
- Antiparticles, 596
- Aperture, 425
- Apex of mirror, 425
- Arc, electric, 259
- Armature, 374
- Asteroids, 619
- Astrocompass, 521
- Atom, 36
- Audibility, threshold of, 355
- Aurora polaris, 310
- Autoclaves, 172
- Avogadro number, 54
- Axis, principal optical, 415
  - secondary optical, 416
- Balance, torsion, 175
- Balmer series, 525
- Base, transistor, 282
- Batteries, solar, 518
  - storage, 216, 253
- Beats, 357
- Belts, radiation, 310
- Bending, lateral, 145
- Beta-rays, 170
- Bias, cut-off, 268
- Bimetal, 166
- Binoculars, 447
- Black body, 403, 488
  - emittance of, 489
- Black holes, 626
- Bohr's atom model, 526
- Boilers, 112
- Boltzmann constant, 67
- Bond, covalent, 142
  - ionic, 142
  - van der Waals, 143
- Boyle's law, 75
- Breakdown, electric in dielectrics, 199
  - in gas, 258
- Breaking point, 150
- Bremsstrahlung, 503
- Brittleness, 124
- Brownian motion, 49
- Bubbles, 110, 130
- Bubble chamber, 566
- Buckling, 145
- Calorie, 82
- Calorimeter, 92
- Camera, photo, 433
- Capacitance, 201
- Capacitors, 204
  - charge of, 204
  - discharge of, 204
  - mosaic, 523
- Capacity, of accumulator, 254
  - thermal, 88
- Capillarity, 125, 131
- Capillary, 131
- Carrier wave, 387
- Cascade showers, 587
- Cathode, 238
  - diode, 266
  - heater, 267
  - reaction, 246
- Cavitation, 124
- Cells, combinations of, 227
  - galvanic, 216, 251
  - granulation, 99
  - Leclanché, 253
  - voltaic, 216
- Celsius, degree, 65
- Centre, optical, 416
- Chain reaction, nuclear, 602
- Chamber, bubble, 566
  - Wilson cloud, 565
- Characteristic, anode, 267
  - current-voltage, 217
  - grid, 268
  - volt-ampere of gas discharge, 256
- Charge, bound, 198
- Charge, electric, 168
  - elementary, 174, 209
  - point, 174
  - polarization, 198
  - space, 266
  - surface, 184
- Chromosphere, 99
- Cherenkov radiation, 398
- Circuit, oscillatory, 374
  - oscillatory, closed, 381
  - oscillatory, open, 139
- Cleavage, 162
- Coefficient, of gas pressure, 63
  - of linear expansion, 135
  - temperature of resistance, 221, 367
  - of viscosity, 135
  - of volume expansion, 163
- Coil, induction, 371
- Collector rings, 364
- Collector, transistor, 282
- Colouring, 479
- Colours, complementary, 479
  - primary, 479
  - spectral, 479
- Comets, 620
- Commutator of generator, 365
- Compressibility of liquid, 124
- Condensation, 101
- Condensation centres, 120
- Conductance, electric, 218
  - specific, 221
- Conductivity, 221
  - acceptor, 275
  - donor, 275
  - n-type, 275
  - p-type, 275
- Conductors, 191, 271
- Connection, parallel, 206, 225
  - in series, 206, 223
- Constant, Boltzmann, 67, 71
  - dielectric, 176
  - electric, 178
  - magnetic, 290
  - Planck's, 397
  - Rydberg's, 525
  - solar, 508
  - Stefan-Boltzmann, 491
  - universal of gas, 70
  - Wiens, 492
- Constellation, 24
- Contact angle, 129
  - potential, 239
- Contraction, linear, 162
  - longitudinal, 144
  - volumetric, 145
- Convection, 83
- Cornea, 434
- Corona, solar, 99
- Cosmogony, 618
- Cosmology, 618
- Coulomb (unit), 177
- Counter, Geiger-Müller, 564
- Crystal, ionic, 142
  - metallic, 144
  - molecular, 143
  - single, 138
  - valence, 142
- Crystal lattice, 138
- Crystallization, 151
- Curie (unit), 570
- Curie point, 303
- Current, alternating, 213
  - direct, 213
  - electric, 210
  - forward, 279
  - reverse, 279
- Defectoscopy, 499
- Defects, crystal, 141
- Deformation, 144
  - absolute, 145
  - plastic, 148
  - relative, 145
- Density, changes in, 153
  - of substance, 33
  - optic, 401
- Deuterium, 578
- Deuteron, 578
- Dew point, 120
- Dewar vessel, 117
- Diamagnetics, 299
- Diameter, effective of molecule, 56
- Diascope, 432
- Dichroism, 463
- Dielectrics, 195, 271
- Dielectric constant, 176
- Diffraction of light, 455
- Diffusion, 39
- Dimension, 27
- Diode, semiconductor, 278
  - vacuum, 266
- Diopter, 417

- Dipole, electric, 196  
   oscillatory, 384  
 Discharge, electric, 257  
   quiet, 257  
   semi-self-maintained, 257  
 Dislocation, edge, 141  
 Dispersion of light, 475  
   normal, 476  
   by prism, 476  
 Displacement, 329  
 Dissociation, electric, 244  
 Domain, 199  
   electric, 200  
   magnetic, 300  
 Donor, 275  
 Doppler effect, 496  
 Ductility, 148  
 Dwarfs, white, 625  
  
 Echo, 353  
 Echo sounder, 361  
 Eddy current, 319  
 Effect, diamagnetic, 317  
   Doppler, 496  
   greenhouse, 121  
   Peltier, 242  
   photoconductive, 516  
   photoelectric external, 509, 511  
   photoelectric internal, 315  
   piezoelectric, 200  
   piezoelectric reverse, 201  
   skin, 378  
   of transformer, 370  
 Efficiency, of heater, 90  
   luminous, 469  
 Einstein's mass-energy formula, 559  
 Elasticity, 147  
 Electricity, quantity of, 169  
 Electrification, 168  
   contact, 174  
 Electrodes, 244  
 Electroextraction, 250  
 Electroluminescence, 531  
 Electrolysis, 244  
 Electrolyte, 244  
 Electromagnet, 313  
 Electrometer, 195  
 Electron, 170  
 Electron-volt, 528  
 Electroplating, 250  
 Electropolishing, 250  
 Electroscopes, 178  
 Element, chemical, 37  
 Elevation, 23  
 Emission, secondary, 259  
   thermionic, 237, 259  
 Emittance, 488  
 Emitter of transistor, 282  
 Energy, 19  
   binding, 581, 583  
   chemical, 44  
   free, 125  
   internal, 44, 80  
   internal of gas, 76  
   kinetic, 44  
   of magnetic field, 322  
   of oscillating body, 338  
   zero-point, 161  
 Envelope of pulse, 386  
 Equation, for harmonic oscillations, 333  
   photo-electric, 513  
   principal of kinetic theory of gases, 60  
 Equilibrium, dynamic, 106  
   thermal, 51  
 Equivalent, chemical, 248  
   electro-chemical, 248  
  
 Epidiascope, 433  
 Episcopes, 432  
 Errors, absolute, 30  
   final absolute, 30  
   random, 30  
   relative, 33  
 Ether, universal, 395  
 Evaporation, 101  
 Events, 45  
   incompatible, 46  
   probability of, 45  
 Expansion, apparent, 165  
   isothermal, 108  
   linear, 162  
   linear coefficient of, 162  
   thermal, 161  
   volume, 74, 163  
   volume coefficient of, 63  
 Extension, longitudinal, 144  
   volumetric, 145  
 Eye, 429  
 Eyeball, 434  
  
 Farad, 201  
   per metre, 177  
 Faraday (unit), 249  
 Faraday's law, first, 247  
   second, 248  
 Ferroelectrics, 199  
 Ferromagnetics, 300  
 Field, electric, 180  
   electromagnetic, 379  
   force, 181  
   gravitational, 181  
   magnetic, 283  
   potential, 186  
 Films, thin, 449  
   wedge-shaped, 432  
 Filters, light, 480  
 Fission, nuclear, 601  
 Flares, chromosphere, 99, 309  
 Fluidity, 124  
 Fluorescence, 532  
 Flux, luminous, 466  
   magnetic, 295  
   of radiation energy, 464  
 Flux linkage, 310  
 Focus, 428  
   principal, 416  
 Force, Ampere, 200  
   elastic, 147  
   electric, 168  
   internal, 146  
   Lorentz, 305  
   magnetic, 288  
   nuclear, 579  
   of internal friction, 134  
   of molecular interaction, 40  
   of repulsion, 48  
   of resistance, 133  
   restoring, 326, 338  
   surface tension, 127  
   thermo-electromotive, 240  
 Foucault, current, 337  
   experiment, 320  
 Fraunhofer lines, 493  
 Freezing, 134  
 Frequency, angular, 327  
   angular resonance, 376  
   cyclic, 332  
   resonance, 352  
 Front, wave, 343  
 Fuel, conventional, 89  
 Function, distribution of speeds, 52  
   Maxwell's, 52  
 Fuses, 236  
 Fusion, 151  
   temperature of, 132  
  
 thermonuclear, 609  
   specific heat of, 152  
  
 Galaxy, 622  
   exploding, 629  
 Galilean transformation, direct, 539  
   inverse, 539  
 Galvanometer, 214  
 Galvanoplastics, 250  
 Gamma-rays, 170  
 Gas, combined law, 69  
   electron, 144, 220  
   ionization of, 255  
   thermal properties of, 68  
   universal constant, 70  
 Gauge, pressure, 57  
 Generation, electron-hole pair, 274  
 Generator, induction, 364  
 Glass, magnifying, 440  
 Gradient, potential, 191  
   velocity, 134  
 Grating, diffraction, 457  
   reflecting, 457  
 Greenhouse effect, 121  
 Grid, of cathode-ray tube, 270  
   of triode, 268  
  
 Half-life, 568  
 Heat, of combustion, 88  
   of condensation, 104  
   mechanical equivalent of, 91  
   quantity of, 82, 86  
   specific, 87  
   specific of fusion, 152  
   of vapourization, 104  
 Heat exchange, 39, 81  
   law, 90  
 Heater, 90  
 Heavy water, 578  
 Helium I, II, 160  
 Henry (unit), 311  
 Hertz (unit), 328  
 Hole, 273  
 Hooke's law, 149  
 Hubble's law, 630  
 Humidity, 118  
   absolute, 119  
   relative, 119  
 Huygens' principle, 398  
 Hygrometer, condensation, 120  
   hair, 120  
 Hyperons, 591  
 Hypsometry, 112  
 Hysteresis, ferroelectric, 199  
   ferromagnetic, 302  
  
 Iconoscope, 522  
 Illuminance, 469  
   first law of, 472  
   second law of, 473  
 Illusion, optical, 439  
 Image, formation of, 427  
   mirror, 406  
   negative, 434  
   positive, 434  
   virtual, 405  
 Imperfections, crystal, 141  
 Index, absolute, 382, 409, 411  
   refractive, 401  
   relative, 408  
 Inductance, 311  
 Induction, electromagnetic, 312  
   electrostatic, 192  
   magnetic, 292  
 Intensity, luminous, 468  
 Interaction, molecular, 47  
   sphere of, 47



- Interference, of light, 447, 454  
 of sound waves, 357  
 of waves, 348  
 Interstitial, 141  
 Ionization, 172  
 impact, 257  
 of gas, 255  
 potential of, 258  
 Ions, negative, 172, 245  
 positive, 172  
 Isochronism, 335  
 Isotherm, 75  
 Isotopes, 53, 172, 568
- Joule (unit), 82  
 Joule's law, 234
- Kaons, 591  
 Kirchhoff's law, 488
- Lamps, daylight, 262  
 incandescent, 236, 469  
 Lantern, projection, 432  
 Laser, 534  
 ruby, 535  
 gaseous, 535  
 Lattice, crystal, 138  
 Law, Boyle's, 74  
 Coulomb's, 175  
 Dalton's, 109  
 Faraday's first, 247  
 Faraday's second, 248  
 Gay-Lussac's, 73  
 gas, 69  
 Hooke's, 149  
 Hubble's, 630  
 ideal gas, 71  
 Joule's, 234  
 Kirchhoff's, 488  
 Newton's of fluid friction, 135  
 Newton's second, 27  
 of charge conservation, 169  
 of conservation of energy, 91, 93  
 of heat exchange, 90  
 of illuminance, 472  
 of photoelectric effect, 511  
 of thermodynamics, first, 95  
 Ohm's, 472  
 Planck's, 492  
 radioactive displacement, 569  
 Stefan-Boltzmann, 491  
 velocity-composition, 540  
 velocity-composition in relativity, 556  
 Wien's, 492
- Layer, surface, 128  
 Leclanché cell, 253  
 Lebedev's experiments, 506  
 Length, proper, 551  
 reduced, 337  
 Lens, 414  
 converging, 416  
 crystalline, 435  
 diverging, 416  
 formula, 419  
 power, 417  
 Leyden jar, 384  
 Light, diffraction of, 455  
 dispersion of, 475  
 interference of, 447  
 polarization of, 461, 463  
 pressure of, 505  
 velocity of, 380, 400  
 Limit, elastic, 150  
 Liquefaction of gases, 117  
 Liquid, 122  
 Long-order, 123  
 Loop, current, 286  
 hysteresis, 199, 302
- Lorentz, force, 305  
 transformations, 547  
 Loschmidt number, 60  
 Luminance, 471  
 Luminescence, cathode, 531  
 Luminophor, 262, 533  
 Luxmeter, 475  
 Lyman series, 525
- Machine, gas-expansion, 117  
 Magnetostriction, 303  
 Magnets, permanent, 284  
 Magnification, lateral, 424  
 of optical instruments, 440  
 Manometer, 58  
 Mass, atomic, 53  
 critical, 604  
 molecular, 53  
 proton, 171  
 relative atomic, 54  
 spectrometer, 576  
 Mass defect, 582  
 Matter, 19  
 Maxwell (unit), 179  
 Melting, 151  
 point, 155  
 Mendeleev Periodic Table, 37  
 Meniscus, 129  
 Meson, 592  
 Mesons,  $K$ , 591  
 $\mu$ , 587  
 $\pi$ , 555, 581  
 Microscope, 442  
 Milky Way, 622  
 Millikan's experiment, 209  
 Mirrors, converging, 425  
 magnetic, 307, 612  
 plane, 405  
 spherical, 425  
 Model, nuclear planetary, 171  
 Moderator, 606  
 Modulation, amplitude, 387  
 Modulator, 387  
 Modulus, elasticity, 149  
 Young's, 149  
 Mole, 54  
 Molecule, 36  
 Moment, electric dipole, 196  
 magnetic, 293  
 Momentum, 561  
 Mosaic, 138  
 Motion, Brownian, 49  
 laminary, 134  
 streamlined, 134  
 Muons, 587
- Neutralization of charges, 169  
 Neutrino, 589  
 muonic, 592  
 Neutrons, 573  
 thermal, 574  
 Newton's law of fluid friction, 135  
 rings, 453  
 second law, 27  
 Nodes, 348  
 Nucleus, atomic, 171  
 Number, atomic, 171  
 Avogadro, 54  
 Loschmidt, 60  
 mass, 574
- Ohm's law, 218, 226, 229  
 Optics, 394  
 geometrical, 401  
 Orbits, atomic allowed, 262  
 Order, long-range, 138  
 Oscillations, elastic, 338  
 forced, 327
- free, 327  
 mechanical, 324  
 natural, 327  
 Oscillator, electron tube, 376  
 Oscillogram, 392  
 Oscilloscope, 390
- Parallax, 24  
 annual, 26  
 measurement of, 25  
 Paramagnetics, 26  
 Parsec, 57  
 Particles, cosmic, 586  
 elementary, 594  
 lambda, 586  
 resonance, 599  
 sigma, 586  
 strange, 586  
 xi, 586  
 Pascal (unit), 55  
 Paschen series, 525  
 Path, mean free, 241  
 Peltier effect, 324  
 Pendulum, 336  
 compound, 337  
 second, 334  
 simple, 334  
 Period, 327  
 of electric oscillations, 375  
 Permeability, 289  
 relative, 289  
 Permittivity, of free space, 176  
 relative, 176  
 Perpetuum mobile, 95  
 Phase, 151  
 diagram, 158  
 difference, 330  
 of oscillations, 329  
 space, 158  
 transitions, 158  
 Photocell, 514  
 barrier, 519  
 Photoconductive effect, 516  
 Photocurrent, saturation, 511  
 Photoelectric effect, external, 509  
 internal, 515  
 laws of, 511  
 Photography, 509  
 Photoluminescence, 531  
 Photometer, 474  
 Photometry, 464, 468  
 Photons, 396  
 Photoresistance, 516  
 Photosynthesis, 508  
 Phosphor crystal, 532  
 Phosphorescence, 532  
 Physics, 19  
 Piezoelectric effect, direct, 200  
 reverse, 201  
 Pi-meson, 555, 581  
 Pions, 581  
 Pitch of sound, 356  
 Planck's constant, 397  
 Planets, 619  
 Plasma, 265  
 Plasticity, 148  
 Plate, deflection, 271  
 Plate, of capacitor, 205  
 of tube, 238  
 P-n junction, 277  
 cells, 518  
 Polarizability, 196  
 Polarization, electronic, 197  
 dipole, 197  
 ionic, 197  
 of cells, 252  
 of light, 461  
 of waves, 460

- Polarizer, 461  
 Polaroid, 463  
 Pole, magnetic, 284  
 Polycrystals, 140  
 Positron, 588  
 Potential, 186  
   electric, 187  
   contact, 239  
 Power, average of a.c. dissipation, 369  
   electric, 223  
   lens, 417  
   resolving, 459  
 Pressure, 57  
   calculation of, 58  
   coefficient of gas, 63  
   critical, 115  
   Laplace, 130  
 Pressure, molecular, 125  
   of light, 505  
 Prism, 412  
   total reflection, 414  
 Probability, 46  
 Process, adiabatic, 97  
   cyclic or closed, 69  
   equilibrium, 69  
   isobaric, 73  
   isochoric, 73  
   isothermic, 74  
 Prominences, 309  
 Proton, 171  
   mass of, 171  
 Pulsars, 626  
 Pupil, 435  
 Psychrometer, 121  
 Pyrometry, 492  
  
 Quantity, physical, 21  
 Quantum, 263  
   theory, 396  
 Quarks, 599  
 Quasars, 629  
  
 Radar, 390  
 Radiation, Cherenkov, 398, 572  
   cosmic, 503  
   gamma, 501  
   induced, 554  
   monochromatic, 448  
   relic, 631  
   stimulated, 534  
   synchrotron, 503  
   thermal, 83  
 Radioactivity, 567  
   natural, 569  
 Radioactive displacement law, 569  
 Radio, 385  
 Radiogalaxies, 503, 623  
 Radioisotopes, 612  
 Radioteleggraphy, 385  
 Radium, 567  
 Radius of molecular action, 42  
 Radon, 567  
 Rainbow, 477  
 Rays,  $\alpha$ ,  $\beta$ ,  $\gamma$ , 567  
   canal, positive, 265  
   cathode, 263  
   cosmic, 585  
 Reactance, capacitative, 368  
   inductive, 368  
 Reactor, nuclear, slow neutron, 605  
   fast neutron, 607  
 Receiver, vacuum tube, 389  
 Recombination, electron hole, 274  
   of ions, 256  
 Rectifier, 267  
   semiconductor, 270  
 Red shift, 623  
  
 Reflection, angle of, 346  
   diffuse, 404  
   factor, 402  
   laws of, 403  
 Reflection, mirror, 405  
   of sound, 359  
   of waves, 346  
   regular, 404  
   total, 411  
 Refraction, angle of, 401  
   laws of, 408  
 Relativity, principle of, 537  
   special, 536  
 Residence, time of, 123  
 Resistance, electric, 218  
   forces of, 133  
   internal, 226  
   specific, 220  
   temperature coefficient of, 221, 367  
 Resistivity, 220  
 Resonance, electric, 383  
   mechanical, 351  
 Resonators, 351, 360  
 Rest, energy, 558  
   mass, 560  
 Retina, 435  
 Reverberation, 359  
 Roentgen (unit), 571  
 Rotor of generator, 364  
 Rule, left-hand, 291  
   right-hand, 314  
 Rutherford (unit), 570  
 Rydberg's constant, 525  
  
 Safety factor, 150  
 Saturation, current, 257  
   magnetic, 300  
 Sclera, 434  
 Screw law, right-hand, 286  
 Second, 26, 530  
 Self-induction, 321  
 Semiconductors, 271  
   extrinsic (or impurity), 275  
   intrinsic (or pure), 273  
 Sensation level, 355  
 Sensitivity, spectral, 467  
 Shear, 145  
 Shielding, electrostatic, 194  
   magnetic, 303  
 Short-circuit, 235  
 Short-order, 123  
 Site, lattice, 138  
 Skin effect, 378  
 Solar constant, 508  
 Solenoid, 286  
 Solid, 137  
 Solidification, 151  
 Solution, 156  
   saturated, 156  
   heat of, 156  
   solid, 157  
 Sound, 353  
   absorption of, 359  
   intensity of, 354  
   loudness of, 354  
   reflection of, 359  
   velocity of, 354  
 Sources, coherent, 349  
   point, 465  
 Spectrograph, 485  
 Spectroscope, 483  
 Spectrum, absorption, 486  
   band, 486  
   continuous, 485  
   diffraction, 477  
   electromagnetic, 502  
   emission, 485  
   infrared, 484  
   line, 486  
   normal, 459, 477  
 Spectrum of radiation, 459  
   solar, 493  
   stellar, 493  
   ultraviolet, 486  
 Speed, molecular, 50  
   root-mean-square, 43, 72  
 Sphere of interaction, 47  
 Spintharoscope, 570  
 Standard, 22  
 Stars, neutron, 625  
 State, of aggregation, 85  
   critical, 114, 116  
   excited of atom, 527  
   gaseous, 47  
   ground of atom, 527  
   of matter, 85  
 Statfarad, 202  
 Stator of generator, 364  
 Steam, dry, 114  
   high-pressure, 114  
   super-heated, 113  
 Stefan-Boltzmann, constant, 491  
   law, 491  
 Steradian, 466  
 Stokes' rule, 531  
 Storms, magnetic, 309  
 Strength, lines of, 182  
   of electric field, 182  
   of magnetic field, 297  
 Stress, normal, 147  
   shearing, 147  
 Structure, long-order, 123  
   short-order, 123  
 Sublimation, 157  
 Sun, 98  
 Sunspots, 98  
 Superconductivity, 222  
 Superfluidity, 161  
 Supernova, 623  
 Superposition of waves, 349  
 System, closed, 45  
   conservative, 93  
   of units, Gaussian, 178  
   of units, SI, 177  
   optical, 428  
   planar, 632  
   spherical, 632  
   solar, 620  
  
 Telescope, astronomical, 444  
   Galileo's, 446  
   Kepler's, 444  
   reflecting, 444  
   refracting, 444  
 Television, 522  
 Temperature, 38  
   of fusion, 152  
   unit of, 65  
 Temperature scale, practical, 65  
   thermodynamical, 66  
 Tension, coefficient of, 126  
   force of, 127  
   surface, 125  
 Thermistors, 275  
 Thermocouple, 241  
 Thermodynamics, first law of, 95  
 Thermoelectricity, 242  
 Thermometer, 38  
   alcohol, 65  
   mercury, 65  
   resistance, 222  
 Thermopile, 243  
 Thomson formula, 376  
 Timbre, 356  
 Time base, 391

- Time, proper, 553
- of residence, 123
- Tokamak, 612
- Tone, harmonic, 357
- fundamental, 357
- Tracers, radioactive, 614
- Transformation ratio, 370
- Transformer, step-down, 369
- step-up, 370
- Transistor, 280
- Transmutations, 568
- man-made, 572
- Triode, 267
- Triple point, 159
- Tritium, 578
- Tsunami, 362
- Tube, cathode-ray, 270
- Tube, television, 524
- Tube, X-ray, 498
- Twisting, 145
- Ultrasound, 360
- Units, of atomic mass, 53
- base, 27
- of density, 29
- derived, 27
- of temperature, 65
- of time, 26
- systems of, 29
- Universe, 618
- Vacancy, 141
- Vacuum, 60
- Value of alternating current, effective, 366
- peak, 366
- van der Waals bond, 143
- Vapour, properties of, 101
- saturated, 105, 107
- unsaturated, 105, 108
- Vapourization, 105
- heat of, 103
- Velocity, of light, 380
- of light in vacuum, 399
- of light in medium, 400
- phase, 344
- of sound, 349
- Vibrations, 345
- Vibrator, 351
- View, angle of, 437
- Viscometer, 136
- Viscosity, 133
- coefficient of, 135
- Visibility curve, relative, 467
- factor, 467
- Vision, persistence of, 437
- stroboscopic, 436
- Vitreous substance, 137
- Volt, 189
- Voltmeter, 214
- electrostatic, 195
- Volume, critical, 116
- Watt, 233
- Watt-hour, 233
- Waves, 334
- coherent, 349
- electromagnetic, 381
- longitudinal, 341
- transverse, 341
- spherical, 343
- standing, 348
- Weber (unit), 311
- Well, potential, 43
- Wetting, 128
- Wilson cloud chamber, 565
- Wind, solar, 100, 627
- Windings, transformer, 369
- Work, 77
- negative, 78
- positive, 78
- total of a current, 232
- Work function, 238
- X-rays, 498
- Zero-point energy, 161